Presented to the Interdisciplinary Studies Program:

UNIVERSITY OF OREGON
APPLIED INFORMATION MANAGEMENT

Applied Information Management

and the Graduate School of the

University of Oregon

in partial fulfillment of the

requirement for the degree of

Master of Science

What CIOs and CTOs Need to Know About Big Data and Data Intensive Computing

CAPSTONE REPORT

Robert P. Brehm

Software Developer

Zoom Software Solutions

University of Oregon
Applied Information
Management
Program

July 2012

Continuing Education
1277 University of Oregon
Eugene, OR 97403-1277
(800) 824-2714

Approved by

Dr. Linda F. Ettinger

Senior Academic Director, AIM Program

Running head: BIG DATA FOR THE CIO AND THE CTO

What CIOs and CTOs Need to Know About
Big Data and Data-Intensive Computing

Robert P. Brehm

Zoom Software Solutions

Abstract

The nature of business computing is changing due to the proliferation of massive data sets referred to as *big data,* that can be used to produce business analytics (Borkar, Carey, & Li, 2012). This annotated bibliography presents literature published between 2000 and 2012. It provides information to CIOs and CTOs about big data by: (a) identifying business examples, (b) describing the relationship to data-intensive computing, (c) exploring opportunities and limitations, and (d) identifying cost factors.

Keywords: big data, data-intensive computing, big data opportunities, big data limitations, big data costs.

Table of Contents

Introduction to the Literature Review

Purpose

Businesses are building massive database systems to drive significant new growth in their business operations (Manyika et al., 2011). These massive databases are identified as *big data* – a phenomenon where the amount of data that a business can collect exceeds its available storage space (Cukier, 2010). Firestone (2010) describes the larger context that has given rise to big data as “the explosion of mobile networks, cloud computing, and new technology [that] has given rise to incomprehensibly large worlds of information” (p. vii). In the Forward to a publication titled The Fourth Paradigm (Bell, 2009), Gordon Bell describes the uses of big data when he examines the phenomenon of new type data-intensive computing in which data sets are being mined for rules and relationships.

Data-intensive computing is a branch of computer science that is distinguished by a need to handle and manipulate large data sets (Cannataro, Talia & Srimani, 2002; Moore, Baru, Marciano, Rajasekar & Wan, 1999). According to Bell (2009), during the last several decades the availability of powerful computer hardware has enabled the creation of data-intensive computing where very large data sets could be exploited. Gray (2007, as cited in Hey, Tansley, & Tolle, 2009) describes data-intensive computing during the last several decades as the *third paradigm* in which data and the software that manipulates the data are required for data reproducibility. Gray and Szalay (2007) further describe how data-intensive computing has transformed into the *fourth paradigm*. Data-intensive computing in the fourth paradigm involves working with ever expanding data sets derived from: (a) large reference data sets, (b) computer algorithms directly applied to data, and (c) automated data collection (Hey et al., 2009). Fourth-paradigm data-intensive computing is manifesting itself as big data (Szalay, 2010).

Many businesses are in a position to deploy big data (Manyika et al., 2011). However, according to LaValle, Lesser, Shockley, Hopkins, and Kruschwitz (2010), existing information technology infrastructures are not going to be replaced with big data systems. As noted by LaValle et al. (2010), big data needs to be integrated with existing systems and business strategy in order to be the most use to business; this places a tremendous burden on businesses as big data is a very new undertaking. Yet Bantleman (2012, April 16) points out that big data does not integrate well into existing IT resources; this requires new spending for big data specific resources and integration. The need for businesses to exploit big data for business advantage is very real, and the businesses that do not successfully embrace big data are going to find themselves at a business disadvantage (Hopkins, 2010).

The purpose for this scholarly annotated bibliography is to identify literature that examines the current practice and implementation of big data systems (Manyika et al., 2011). The goal is to introduce the concepts of (a) fourth-paradigm data-intensive computing (Kouzes, Anderson, Elbert, Gorton, & Gracio, 2009) and (b) big data (Manyika et al., 2011) to Chief Information Officers (CIO) or Chief Technology Officers (CTO) so that they may better understand (a) why and (b) how to integrate and deploy big data.

Problem

Data-intensive computing – computing performed on large data sets – has been in practice for decades (Gray & Szalay, 2007). Initially data-intensive systems were built with expensive mainframe technology but the efficiencies were poor (Joshi, 2005). Several major advances in computing such as grid computing (Foster, 2002) – using inexpensive hardware to process data in parallel – and high speed networks (Johnston, 1998, July) dramatically increased the capabilities of data-intensive computing for computational simulation which Gray (Gray, 2007, as cited in Hey et al., 2009) describes as the third paradigm.

The emergence of large reference data sets and massively growing raw data streams caused data-intensive computing to become big data (Bell, 2009). Big data is analyzed by data scientists into understandable patterns (Eagle, 2010) rather than by simulation. From these resultant patterns, actions are taken – either via interaction with pattern visualization (i.e. heat map analysis) or via automation (i.e. automatic data processing) (Kenwright, 1999). This is the essence of Gray's fourth paradigm (Gray & Szalay, 2007).

Big data systems are being implemented in multiple enterprise sectors, including commerce, science, and society (Bryant, Katz, & Lazowska, 2008). A few examples are provided to illustrate the use of big data in real world settings. These examples share several common traits: (a) they utilize large data stores, (b) they apply domain appropriate analysis, and (c) they present the analytic results visually (Manyika et al., 2011).

Example 1: Stock market. The stock market is increasingly relying on program trading (algorithmic trading or high-frequency trading) to derive significant advantages (Pallay, 2005). Program trading has become so prevalent that non-exchange trading facilities have been developed to settle accounts from program trading (Carrie, 2006). Even Twitter feeds have been mined as program traders have discovered that twitter feeds are correlated to stocks prices (Ruiz, Castillo, Hristides, Gionis & Jaimes, 2012).

Example 2: Public health. An alert system was put into place by the Global Public Health Intelligence Network to monitor for outbreaks of a mysterious respiratory disease by utilizing the HealthMap system (Brownstein, Freifeld & Madoff, 2009). HealthMap is a big data system that takes in data from diverse sources such as online media reports and government alerts, applies text processing the alerts, classifies the alerts, and then overlays the alerts on a map (Freifeld, Mandl, Reis & Brownstein, 2008).

Example 3: ATDS. An enhancement to the Advanced Terrorist Detection System was proposed that actively mines the web for web user’s activity and compares it with activity in proximity to the user's location (Shapira, Elovici, Last, & Kandel, 2008). Any activity above a threshold generates an alarm to counter-terrorist organizations. While it is impossible to say if this system was ever implemented for security reasons it is certainly within technological feasibility.

Example 4: Sloan Digital Sky Survey (SDSS). This ambitious astronomy project has successfully mapped over ¼ of the sky, and has taken pictures of 300 million celestial objects (Hey, n.d.). A dedicated telescope equipped with CCD cameras is combined with a special software pipeline to automatically collect celestial photographs and data (Sloan Digital Sky Survey, 2012). An affiliated project – the SkyServer – makes SDSS available to the public (Hey, n.d.).

Example 5: Commercial system. Walmart has built a private big data system to support its retail operations. The Walmart big data system contains 140 terabytes and grows with over 1 million customer transactions per hour. Walmart’s CIO, Rollin Ford states “every day I wake up and ask, how can I flow data better, manage data better, analyze data better?” (as cited in Cukier, 2010).

According to Bantleman (2012, April 16) the implementation of big data requires a high level of sophistication. He reports that a new expert known as the *data scientist* has emerged; and when these individuals are not available businesses incur costs to retrain their workforce for big data. Further, he reports that these systems are built on technology that is relatively new. As an example, Bantleman (2012, April 16) cites that traditional databases are not well suited for big data; new data storage technology needs to be used. Finally, he notes that major costs are incurred in integrating big data into the existing IT environment.

Research Questions

Main question. What are the key factors that a CIO or CTO needs to consider when integrating big data into the business?

Sub questions.

- What is data-intensive computing and how is it related to big data (Hey, 2010)?
- What are the current opportunities of big data system applications (Bryant et al., 2008)?
- What are the current limitations of big data system applications (Jacobs, 2009)?
- How are costs currently estimated for implementation of big data systems (Trelles, Prins, Snir & Jansen, 2011)?

Significance

Companies are rushing to incorporate big data systems into their existing infrastructure due to the proliferation of inexpensive big data components (Bantleman, 2012, April 16). Manyika et al. (2011) predict that gains from the use of big data systems will be the greatest in certain sectors such as computer products, finance, insurance and government while the manufacturing and health care industries will see fewer gains initially from the use of big data. As explained by Manyika et al. (2011), the health care industry has typically suffered from data silos –data residing in pools that are not connected to one another. Connecting the health industry data pools will reap tremendous advances in the health care industry (Manyika et al., 2011).

Manufacturing is tied to productivity gains, and big data increases productivity across the value chain by integrating data sources (Manyika et al., 2011). Also, the manufacturing sector may reap benefits from deploying big-data systems for real-time decision making (Ali, Chan & Lee, 2008) in smart factories. As the number of sensors such as RFID tags increase rapidly, big data will result and all systems in a manufacturing environment – design, factory floor, engineering and management - will harvest and utilize this data (Manyika et al., 2011).

Bryant et al. (2008) warns that funding for big data research has lagged behind the business adoption of big data. Gray and Szalay (2007) warn that big data requires new tools for information management and data flow. Yet the CIO and CTO operate in an environment of uncertainty regarding big data. For example, the federal government provides incentives for the *meaningful use* of electronic health records (EHRs) as set forth in the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009; this Act that provides incentives for health care data mining and reporting of quality metrics to the government (Jha, 2010). Meaningful use is expanded in Stage 2 of the EHR Incentive Program to include mandates for the sharing of real time, high volume data between health entities and penalties for health entities that do not achieve meaningful use goals by 2015 (Copoulos, 2012).

Audience

The audience for this annotated bibliography is the CIO and the CTO. These key business executives are responsible for the alignment of big data with other business objectives (LaValle et al., 2010). For big data to be successful in a business LaValle et al. (2010) advise that information managers must (a) link big data with business strategy, (b) provide an easy to understand end user experience, and (c) be linked to existing key business processes so that action can be taken at the right time.

Companies are rushing to incorporate big data systems into their existing infrastructure due to the proliferation of inexpensive big data components (Bantleman, 2012, April 16). Yet Bryant et al. (2008) warn that funding for big data research has lagged behind business adoption of big data. The CIO and CTO operate in an environment of uncertainty regarding big data. Earl and Scott (1999) describe the emergence of a new executive in some organizations – the Chief Knowledge Officer (CKO) – to deal with the special requirements of knowledge flow.

Delimitations

Topic definition. Data-intensive computing has moved from the third paradigm to the fourth paradigm, and the transition is manifesting big data (Szalay, 2010). Business is moving aggressively to implement big data data–intensive computing (Manyika et al., 2011). However, businesses are incurring significant costs to implement big data (Bantleman, 2012, April 16). Based on this perspective, it is clear that that the success of big-data in business is going to be dependent on how well the CIO and the CTO perform in leading the adoption and implementation of big-data in business **(**LaValle et al., 2010).

Time frame. The time frame for articles selected for use in this study is primarily within the last 11 years (2000 – 2012). Doug Laney, Vice President of Research at Gartner Group, first described big data (although he did not use the term big data) in 2001 as "current business conditions and mediums are pushing traditional data management principles to their limits, giving rise to novel, more formalized approaches" (Laney, 2001).

Audience. The audience is limited to the CIO /CTO and other interested executives. For big data to be successful in business, LaValle et al. (2010) advise that information managers must link big data with business strategy, provide an easy to understand end user experience, and be linked to existing key business processes so that action can be taken at the right time.

Focus. Lutchen (2004, pp. 8-13) describes that IT management activities need to be examined through a “lens” composed of six elements: (a) alignment to business goals, (b) resiliency of resources from disruption, (c) integration of emerging technology, (d) support of IT (staffing, finance, communication), (e) operations, and (f) leveraging IT assets across the business. The literature selected for this review focuses on (a) big data alignment to business goals, (b) integration of big data as an emerging technology, and (c) leveraging big data in the business.

Literature collection criteria. It is important to have well-established criteria to be able to collect appropriate literature. The following criteria is used for collecting the literature:

- Use online databases such as University of Oregon and Google Scholar (Olhoff, 2011).
- Read professional articles and books, and then look at the reference section to find additional literature (Olhoff, 2011).
- Make a list of the primary keywords. Also consider recording synonyms for the keywords (Olhoff, 2011).
- Start with broad synthesis of literature like those found in Encyclopedias and Wikipedia (Creswell, 2009, p. 32).
- Use articles in established national journals (Creswell, 2009, p. 32).
- Search for major conferences and seek out the papers that were presented (Creswell, 2009, p. 33).

Preview of the Reading and Organization Plan

Reading plan preview. Although there is no one way to conduct a literature review, Creswell (2009) notes that many scholars proceed "in a systematic fashion" (p. 29). Busch et al. (2005) have developed a process based on *conceptual analysis* that is used to guide the deep reading process of the literature selected for analysis in this annotated bibliography. The focus of the analysis is framed by the concepts embedded in the research questions.

Organization plan preview. The results from the reading plan need to be analyzed to answer the research questions presented above. Bastek, Robinson and Barnes (2012) provide four different methods for organizing the results. Of these four methods the State of the Art review is used when the review concentrates on the most current literature in a research area (Bastek et al., 2012). The State of the Art review is appropriate for the organization of information in the Annotated Bibliography section of this paper and for the presentation of the data analysis results in the Conclusions.

Definitions

The subject of big data contains terms and concepts that are often less familiar than terms and concepts of more traditional information technology. In the paper *Big Data Management: Ogres, Onions, or Parfaits?* Borkar et al. (2012) describes the current state of big data management as quite chaotic. The CIO and CTO need to understand the new vocabulary to be effective in managing big data.

Alert System – A system that is used to identify or prevent outcomes based on: (a) an integrated computerized database, (b) an alert-generating program, and (c) a reliable system for alert notification (Raschke et al., 1998, October 21).

Analytics – Same as data mining (Kohavi, Rothleder, & Simoudis, 2002).

Big Data – The precise definition of big data has not been established (Franks, 2012). This definition is offered by Manyika et al. (2011): "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze" (Executive summary, para. 2).

Business Intelligence – Cohen, Dolen, Dunlap, Hellerstein, and Welton (2009) define business intelligence as "software [tools that] produce reports and interactive interfaces that summarize data via basic aggregation functions over various hierarchical breakdowns of the data into groups" (1. Introduction, para. 1).

Chief Information Officer (CIO) – Broadbent and Kitzis (2005) define the CIO as "the most senior executive responsible for identifying information and technology needs and then delivering services to meet those needs" (p. 6).

Chief Technology Officer (CT0) – According to Smith (2003, July-August), businesses responded to the increasing role of emerging technology on business strategy by appointing a

CTO. Smith further states that the CTO is responsible for monitoring new technologies and overseeing the adoption of the new technology into the business among other duties.

Cloud Computing – According to the National Institute of Standards and Technology, cloud computing is defined as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” (Mell & Grance, 2011, 2. The NIST definition of cloud computing, para. 1).

Correlation - Encyclopædia Britannica (Correlation, 2012) defines correlation as “In statistics, the degree of association between two random variables” (para. 1).

Data-Intensive Computing – An early definition of data-intensive computing is computer applications that “devote a large fraction of [computer] execution time to movement of data” (Moore et al., 1999, p. 107). Cannataro et al. (2002) define it as “applications that explore, query, analyze, visualize, and, in general, process very large scale data sets” (para. 2). However, to Kouzes et al. (2009) this definition needs to be broadened from one focusing solely on data sets to “a broader realm of issues dealing with the time to reach a solution” (p. 27). Kouzes et al. (2009) give some examples of the broader issues faced in data-intensive computing: the processing of massive data streams in real time, the gathering and processing of information in revolutionary ways, and the presenting knowledge to the end user.

Data Mining – - Encyclopædia Britannica (Data mining, 2012) defines data mining as “the process of discovering interesting and useful patterns and relationships in large volumes of data” (para. 1).

Data Scientist – Rappa (2012, May 3) defines a data scientist as a person who possesses skills in statistics, applied mathematics and computer science. He further states that "it's fair to say that statisticians would be the closest to what we call a data scientist in the traditional sense" (Data scientist, analyst or statistician, para. 5).

Data Warehouse – A data warehouse is a specialized database that contains archival data from other databases and is used for data reporting and visualization (Information system, 2012).

Distributed Computing - Encyclopædia Britannica (Distributed computing, 2012) defines distributed computing as "a method that researchers use to solve highly complicated problems without having to resort to an expensive supercomputer" where "large numbers of computers [are used] to split up the computational load" (para. 1).

Distributed File System – Silberschatz, Peterson, and Galvin (1991, p. 491) define a distributed file system as a file system that is stored across loosely coupled servers interconnected by a computer network.

e-Manufacturing – Koc, Ni, Lee and Bandyopadhyay (2004) define e-manufacturing as "a transformation system that enables the manufacturing operations to achieve predictive near-zero-downtime performance as well as to synchronize with the business systems through the use of web-enabled [and wireless technologies]" (p. 97-2).

e-Science (eScience) – Hey and Trefethen (2003) provide this definition of e-Science as "the increasingly global collaborations – of people and shared resources – that will be needed to solve the new problems of science and engineering" (1. Introduction, para. 1).

Expert System – Encyclopædia Britannica (Expert system, 2012) defines an expert system as “a computer program that uses artificial intelligence to solve problems within a specialized domain that ordinarily requires human expertise” (para. 1).

Four Paradigms – Gray (2007) describes four paradigms of scientific inquiry extant in modern history (a thousand years ago to the present) as: (a) in the first paradigm (a thousand years ago) scientists used empirical methods, (b) in the second paradigm (a few hundred years ago) scientists used models and generalizations, (c) in the third paradigm (last few decades) scientists used computational simulation, and (d) in the fourth paradigm (today) scientists are unifying theory, experiment, and simulation to explore data for relationships**.**

Fourth Paradigm – The fourth paradigm is defined as the unification of theory, experimentation and simulation that allows data to be explored dynamically (Gray, 2007). This paradigm is evolving today due to two major factors: (a) the emergence of large reference data sets and (b) data sets are growing massively due to data computation and automated collection (Hey, 2009). Bell (2009, March) states that the promise of the fourth paradigm has not been realized fully due to data complexities and a general lack of understanding of the topic.

Hadoop – Olson (2011, January 12) states that Hadoop is a “platform [that] was designed to solve problems where you have a lot of data – perhaps a mixture of complex and structured data – and it doesn’t fit nicely into [conventional database] tables. It’s for situations where you want to run analytics that are deep and computationally extensive” (para. 4).

MapReduce - Warden (2011, MapReduce, para. 1) describes the primary limitation of a tradition database: all processing occurs in a single database instance in a highly structured manner. This is sub-optimal for big data. The goal of MapReduce is to “create a pipeline that reads and writes to arbitrary file formats, with intermediate results being passed between stages as files, with the computation spread across many machines" (MapReduce, para. 1).

Machine learning – Langley (1996) describes machine learning as providing machines with the ability to acquire intelligence through automatic means via "knowledge representation, memory organization, and performance" (p. 1) in a carefully chosen domain.

Metadata – Bargmeyer and Gillman (n.d.) defines metadata as "data used to describe other data so that the usage turns it into metadata" (para. 1).

NoSQL Database – Warden (2011, NOSQL Databases, para. 1) defines a NoSQL database as a type of database where data is stored as name (key)/value pairs instead of a traditional SQL database such Oracle or Access. NoSQL databases are faster and more flexible then SQL databases but require custom programming in order to store and access the data.

Parallel Computing – Barney (n.d.) explains traditional computing as a computational technique where processing occurs in a serial fashion. The problem is that serial computing cannot take advantage of distributed computing. Barney defines parallel computing as "the simultaneous use of multiple compute resources to solve a computational problem: (a) to be run using multiple [processors] (b) a problem is broken into discrete parts that can be solved concurrently, (c) each part is further broken down to a series of instructions, and (d) instructions from each part execute simultaneously on different [processors]" (What is parallel computing, para. 1).

Parallel Database- Sokolinsky (2004) defines a parallel database as one where a database is implemented on a multi-processor system with a high degree of connectivity.

Open Source – Encyclopædia Britannica (Open source, 2012) defines open source as a "social movement, begun by computer programmers, that rejects secrecy and centralized control of creative work in favour of decentralization, transparency, and unrestricted ("open") sharing of information. Source refers to the human-readable source code of computer programs, as opposed to the compiled computer programming language instructions, or object code, that run on computers but cannot be easily understood or modified by people" (para. 1).

Program Trading – Finnerty and Park (1988, Winter) define program trading as "the purchase or sale of a portfolio of securities by institutional investors as if the portfolio were one stock" (p. 40).

Service - Papazoglou and Georgakopoulos (2006) define services as "self-describing, open components that support rapid, low-cost composition of distributed applications" (p. 2, Overview of Services, para. 1) (see Distributed Computing).

Service-oriented Computing – Papazoglou and Georgakopoulos (2006) describe service-oriented computing as a "computing paradigm that utilizes services as the fundamental components for developing applications" (p. 25, para. 1) of layered components.

Sharding – Warden (2011, Sharding, para. 1) describes sharding as a technique to efficiently distribute rows in a database table across multiple machines by using each row's unique identification key.

Third Paradigm – Gray's third paradigm is identified as the emergence of computational simulation in the middle of the 20th century (Bell, 2009) due to advances in computing technology. Bell characterizes the scientific record in the third paradigm as using data summaries rather than full documentation as in the second paradigm. Also, in the third paradigm, Hey et al. (2009) state that the software used to manipulate and simulate the data had become an integral part of experimental reproducibility along with the data**.**

Transaction Processing – Borkar et al. (2012) define transaction processing as a " system that underlies the online application that powers a business' day-to-day activities and is the main producer of the large volume of data that is filling the business' data warehouse" (2.1 big data in the database world, para. 3).

Visualization – According to Friendly (2009), data visualization is "the science of visual representation of "data", defined as information which has been abstracted in some schematic form, including attributes or variables for the units of information" (p. 2).

Research Parameters

The literature selected for examination in this annotated bibliography focuses on big data technology, opportunities, limitations and costs from the CIO and the CTO perspective. The search is broken down into three areas: (a) big data technology and relationships to data-intensive computing, (b) big data opportunities and limitations, and (c) big data economics.

The research parameters section consists of descriptions of (a) the search strategy including search terms and resources, (b) the documentation approach, and (c) the criteria used to evaluate and select literature, and (d) the reading and organization plan.

Search Strategy

Search terms. Search terms are derived during the exploratory search process. Henshen (2011), executive editor of Information Week, provides an article on Hadoop that is an excellent source of keywords. In addition, the *Big Data Glossary* provides corroboration of the keys words identified by Henshen (cited in Warden, 2011). The following search terms are giving good results for big data technology.

- big data
- data-intensive computing
- distributed computing
- grid computing
- network attached storage
- parallel computing
- data visualization
- Hadoop
- cloud computing

- map-reduce
- distributed database
- machine learning
- nosql database
- grid computing
- analytics
- semantic web
- Influential authors in the field of big data:
    - Gordon Bell
    - Doug Cutting
    - Jim Gray
    - Tony Hey
    - Doug Laney

Big data opportunities and limitations are best found by combining technology search terms with the words *opportunities* and *limitations*. So the following are useful search terms:

- big data opportunities
- big data limitations

In addition, these search terms are useful as well:

- Fourth paradigm
- eScience, e-science

Useful search terms for researching big data costs include:

- Big data investment
- Big data costs

- Information technology investment

Search resources. Creswell provides a search priority plan for acquiring review literature. However, Creswell's plan is oriented towards using a physical library rather than online resources. For example, Creswell advises searching through the *Dissertation abstracts* microfiche archives. A version of this dissertation abstract is available online, and the online version is used instead.

The following resource categories are prioritized by Creswell (2009, pp. 32-33):

1. **Encyclopædias**. Creswell recommends using an Encyclopædia for an initial survey of a topic and to find an introductory reference list. *Wikipedia* and *Encyclopædia Britannica online* are used for this purpose.
2. **Journals.** Journal articles in publications that are highly read and cited are considered to be high value "especially those that report research studies" (Creswell, 2009, p. 32). The *MIT Sloan School Management Review* and the *Harvard Business Review* are examples of such journals.
3. **Books**. Creswell recommends "scholarly monographs that summarize scholarly literature" (Creswell, 2009, p. 33) followed by other books or parts of books on a specific topic. *Big data bibliography– a free bibliography from Safari Books Online* (2011) provides a good list of current books in the subject area of big data. *Big data: The next frontier for innovations, competition and productivity* (Manyika et al., 2011) provides a good overview of present state of big data, and it also talks about the future of big data. *Managing IT as a business* (Lutchen, 2004) and *The new CIO leader* (Broadbent & Kitzis, 2005) provide excellent background information on IT in business and the role of the CIO in managing IT, respectively.

4. **Recent conference papers.** Recent papers and possibly presentation slides provide excellent information into the current thinking on a topic. For example, a big data conference was held in Washington, DC on May 8-9, 2012, and the conference agenda can be mined for recent published papers (Big data conference, 2012).
5. **Dissertation abstracts**. Creswell (2009) cautions that dissertation abstracts vary in quality, and any work found in the *Dissertation abstract* online web database needs to be scrutinized.
6. **Web resources**. Search databases that are rich in scholarly sources include Google Scholar, IEEE Computer Science Digital Library, ACM Digital Library, CiteSeer, and WorldCat.org. Search engines used for searching include Google Scholar and to a lesser extent Google. Google articles sometimes yield quality works but articles need to be verified for quality.

In addition to Creswell's priority search plan given above, the following source category is added:

7. **Online trade journals**. Online trade journals are useful for providing quality background information. *CIO Magazine* online edition provides information to top information technology executives. It offers news, information, and analysis on a variety of topics including big data, cloud computing and visualization. *CIO Insight* online edition is a similar publication to CIO magazine but is aligned with real world problems. *Teradata* magazine concerns itself with issues of big data from a managerial perspective.

Literature Evaluation Criteria

Credible resources are defined as those that have been cited, appear in peer-reviewed journals, or appear in influential publications. Bell and Smith (2009) have developed an informative guide for evaluating credibility of references based on: (a) authority, (b) objectivity, (c) quality, (d) coverage, and (e) currency. From this guide the following selection criteria are used:

- Is an author given?
- What are the author's credentials, reputation, and institutional or organizational affiliations?
- Is the work free of bias?
- Is the article reflecting the author's affiliations?
- Is the article well written and referenced?
- Is the work's content verified by other works?
- When was the work published?
- Is the work linked to the search terms?
- Is the work relevant?
- Is the work scholarly as defined in *A Guide to Evaluating Resources* (n.d.)?

Documentation Approach

Machi and McEvoy (2009) advise that "a methodical approach to searching the literature and reflective deliberation on the impact of the literature on your topic will provide a sound foundation to your literature review" (Step 2. Search the literature, Stage 2, para. 1). A sound documentation approach helps in the search and deliberation process by providing the means to organize and relate the acquired literature.

The literature searching approach used for this literature review relies heavily on web base searching. In order to archive web searching results the *Zotero Firefox browser plugin* is used. Zotero provides a simple button that when clicked records a web site URL. Recorded URLs are moved into categories that are created by the user to fit the user's requirements (Zotero, n.d.).

Machi and McEvoy (2009) state:

> One note of caution before leaving this topic [of using electronic resources]: many of the journals provided through university electronic databases connect directly with your personal research databases (such as EndNote, Citation, and Ref Works). This means that you can cite a journal, transfer its abstract, and catalog its contents with one click of your mouse. The good news is that you can document and catalog this information quickly. The bad news is that little, if any, of this knowledge transfers to your consciousness. Make use of the great improvements electronic databases provide to the task of searching, but take the time to understand and internalize the meaning of your information as you collect your data. (Step 2. Search the literature, Stage 4, para. 3)

In order to address this, two actions are taken (a) use Zotero's notes feature to record notes about entries, and (b) employ mind mapping software.

The *Mind map,* invented and popularized by Tony Buzan (Druce, 2010; Mind map, n.d.) is an outlining tool where ideas radiate outward from a central idea (Budd, 2004). Beel, Gipp, and Stiller (2009) recommend using mind maps for document summarization and advanced searches. *Freeplane* (n.d.) is a free mind map tool that is used for this literature review for outlining, summarizing, and advanced searches.

Reading and Organization Plan

The reading and organizational plan is built around the concept of content analysis which is a research tool used to select works based on the presence of "certain words or concepts within texts or sets of texts" (Busch et al., 2005, An Introduction to Content Analysis, para. 1). Busch et al. (2005) further identify two styles of content analysis: (a) Conceptual Analysis – establishing the existence and frequency of concepts in a text, and (b) Relational Analysis – looking for deeper meaning between concepts in a text. For this literature review Conceptual Analysis is used.

Reading plan. According to Busch et al. (2005), conceptual analysis starts with the selected research questions and then applies search phrases to selected texts to determine if the texts are relevant to the research questions. Once the literature is selected, Busch et al. (2005) describe an eight step coding process which is used to guide the development of the reading plan, which is used to identify concepts embedded in the research questions.

1. **Decide on level of analysis**. Are single keywords or phrases appropriate? For this analysis search phrases such as *big data* and *data-intensive computing* are the most relevant. Certain coding terms such as eScience and Analytics are single keywords.
2. **Decide how many concepts to code for.** Are predefined or interactive concepts appropriate? For this analysis predefined concepts are most relevant. Key concepts include (a) big data technology and relationships to data-intensive computing, (b) big data opportunities and limitations, and (c) big data economics.

3. **Decide whether to code for existence or frequency of a concept.** In this review the existence of a concept in the work is the most productive due to the scarcity of quality research material.
4. **Decide on how you will distinguish among concepts.** How much leeway is given if concepts differ in their form? For this review significant leeway is given as the vocabularies are still evolving. As an example, the coding keyword e-Science often appears in works as eScience. Another difficulty is that early works about big data do not directly use the term big data as the term was not in widespread use (Laney, 2001). Therefore, flexibility is required in this coding process.
5. **Develop rules for coding your texts.** How will discrepancies in concepts be dealt with? For this review discrepancies from the concepts in individual texts are noted but included in concept categories. In many cases the context of the work is considered in the coding process. For example, many authors may describe big data without actually using the key phrase, yet the work may be validly included in the coding process. Rules helpful for resolving this situation are (a) is the author describing a big data concept? and (b) do valid references to the work exist that corroborate that this work describes big data?
6. **Decide what to do with irrelevant information.** How will words such as *and* and *the* be handled? For this review they are ignored.
7. **Code the texts.** Coding involves the identification of key concepts in the work that address the research questions. Busch et al. (2005) notes that while automated searches use keywords and key phrases efficiently to find matching words and

phrases, the researcher needs to carefully examine how well words and phrases match the concepts being sought.

8. **Analyze your results.** The researcher needs to decide how to process the data results from the coding process. For this review all data from the coding process is examined for relevance to the research questions and organized in relation to the Organization Plan described below.

Organization plan. Bastek et al. (2012) describe five types of review papers: (a) state of the art review, (b) historical review, (c) a comparison of perspective review, (d) a synthesis of two fields review, and (e) a theoretical model building review. For this study a state of the art review plan is most appropriate as a way to organize the results of the literature analysis (the coding). A state of the art approach refers to an examination of the most current research in big data. The nature of the exploration of big data is to familiarize the CIO and the CTO with the state of the art in big data from a managerial point of view – the technologies, their opportunities, their limitations and their costs. However, a certain amount of historical analysis is necessary for context and completeness. What follows is set of potential state of the art concepts that are organized and presented in relation to each research question.

Main question. What are the key factors that a CIO or CTO needs to consider when integrating big data into the business?

This section is organized around issues that CIO's and CTOs are concerned about including (a) alignment to business goals, (b) integration of emerging technology and (c) leveraging big data in the business (Lutchen, 2004, pp. 8-13).

Sub questions.

- What is data-intensive computing and how is it related to big data?

  Data intensive computing is composed of common components of (a) data acquisition, (b) data management, (c) modeling and simulation, (d) algorithms, (e) information analytics, and (f) computing platforms (Kouzes et al., 2009). Big data is a type of data intensive computing that requires new common components to meet enormous data challenges (Gorton et al., 2008, April).

- What are the current opportunities of big data system applications?

  Big data will be widely adopted in the commercial, science, medical and health, and government sectors (Bryant et al., 2008). Manyika et al. (2011) write that certain economic sectors such computer, electronic products, finance, insurance and government will see the most adoption of big data but eventually all sectors will have to adopt big data.

- What are the current limitations of big data system applications?

  Lynch (2008, September) proposes that big data is constrained by (a) cost, (b) lack of standards for data description and exchange, (c) data preservation. Bantleman (2012, April 16) states that a shortage of data science professionals is constraining the adoption of big data. Jacobs (2009) states that big data reaches limitations imposed by computer hardware, and the implementation of big data requires costly distributed computing.

- How are costs currently estimated for the implementation of big data systems?

  All of the components and personnel required to implement big data need to be carefully considered by business executives and especially the CIO and CTO. Central to these executives’ roles in business is the concept of alignment of IT to business (Lutchen, 2004, pp. 8-10). Bantleman (2012, April 16) warns that the biggest costs in big data are in the integration of big data into the existing IT infrastructure. Trelles et al. (2011) warn that expensive data storage hardware costs are a gating cost.

Annotated Bibliography

The annotated bibliography contains 32 references. The literature is grouped into five main categories that align with the Organization Plan presented in the Research Parameters section of the paper. The KU Writing Center (2011, July) writes that an annotated bibliography entry should contain at least one of the following elements: (a) a descriptive element– a summary of the book or work and how the author addresses the topic, and (b) an evaluative element– a summary of how successful the author is in achieving the objective. In addition, in this annotated bibliography an element examining the credibility of each reference is included, in relation to the literature evaluation criteria provided in the Research Parameters. Each annotation below contains the following information: (a) a citation, (b) an abstract, (c) a descriptive summary, and (d) an evaluation of the author's credibility.

Key Factors to Consider When Integrating Big Data into a Business

This section is organized around issues that CIOs and CTOs are concerned about including (a) alignment to business goals, (b) integration of emerging technology and (c) leveraging big data in the business (Lutchen, 2004, pp. 8-13).

Agrawal, D., Das S., & Abbadi, A. (2011). Big data and cloud computing: Current state and future opportunities. *Proceedings of EDBT*, 530-533. Retrieved June 2, 2012 from http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf.

Abstract. Scalable database management systems (DBMS)—both for update intensive application workloads as well as decision support systems for descriptive and deep analytics—are a critical part of the cloud infrastructure and play an important role in ensuring the smooth transition of applications from the traditional enterprise infrastructures to next generation cloud infrastructures. Though scalable data management has been a vision for more than three decades and much research has focused on large scale data management in traditional enterprise setting, cloud computing brings its own set of novel challenges that must be addressed to ensure the success of data management solutions in the cloud environment. This tutorial presents an organized picture of the challenges faced by application developers and DBMS designers in developing and deploying internet scale applications. Our background study encompasses both classes of systems: (i) for supporting update heavy applications, and (ii) for ad-hoc analytics and decision support. We then focus on providing an in-depth analysis of systems for supporting update intensive web-applications and provide a survey of the state-of-the-art in this domain. We crystallize the design choices made by some successful systems large scale database management systems, analyze the application demands and access patterns, and enumerate the desiderata for a cloud-bound DBMS.

Summary. Cloud computing has become an increasingly important type of service-oriented computing because cloud computing charges users very low upfront costs and provides users the ability to scale the level of cloud computing consumed to fit the required computing load. Data-intensive computing initially focused on parallel databases and distributed computing but has given way to NoSQL databases and Hadoop as the open source choice implementation of MapReduce. The authors focus on factors that make cloud computing attractive such as: (a) scalability, (b) fault-tolerance, (c) and cost. Cloud computing is useful for single large database applications and also applications that have numerous small databases. Two problems that are not fully addressed in big data cloud computing are the limitations of NoSQL databases and the ability to handle sudden processing spikes.

Credibility. Divyakant Agrawal is a professor of Computer Science at the University of California, Santa Barbara. Agrawal is also an ACM distinguished scientist. Sudipto Das is a doctoral candidate in the Department of Computer Science at University of California, Santa Barbara. Amr El Abbadi is currently a Professor and Chair of the Department of Computer Science at the University of California, Santa Barbara. El Abbadi is a fellow of the ACM. The EBT association is a non-profit organization for promoting research in databases and information technology. The proceedings are peer-reviewed. This article is cited by 6 authors.

Bollier, D. (2010). *The promise and peril of big data*. Washington, DC: The Aspen Institute. Retrieved June 11, 2012 from https://www.c3e.info/uploaded_docs/aspenbig_data.pdf.

Abstract. Ever-rising floods of data are being generated by mobile networking, cloud computing and other new technologies. At the same time, continued innovations use advanced correlation techniques to analyze them, and the process and payoff can be both encouraging and alarming. *The Promise and Peril of Big Data* explores the ways these inferential technologies can positively affect medicine, business and government, and also examines the social perils they pose. Written by conference rapporteur David Bollier, the report summarizes the insights of the Eighteenth Annual Roundtable on Information Technology, which sought to understand the implications of the emergence of "Big Data" and new techniques of inferential analysis.

Summary. Bollier's work is introduced by a forward by Firestone (2010), who states the problem (and opportunity) of big data is that the explosion of data derived from sensor networks is overwhelming the means of collecting and understanding the data. Bollier first states that "a radically new kind of "knowledge infrastructure" is materializing. A new era of big data is emerging…." (p. 1). Certain factors, Bollier reports, are contributing to the new era of big data: (a) technology such as computer storage (b) data streaming devices such as video cameras, telescopes and traffic monitors, (c) cloud computing, and (d) consumer-oriented applications such as Google Earth and MapQuest. Bollier explores the question of whether statistical correlations can take the place of scientific modeling. For example, he states that a pitfall of big data is that a correlation made for a group may not apply very well to an individual in the group. Visualization, Bollier states, is an important tool in using big data. However, Bollier warns, visualizing data in a commercial setting is less about truth and understanding and more about making money. Data needs to be selected judiciously to give the desired results; one has to be careful to only collect data that is going to be meaningful. Bollier reports that as big data improves real-time trends are used to more accurately predict future outcomes and to acquire new customers. In a special section on health care Bollier records that big data contributes to health care in two ways: (a) population care and (b) personalized health care. But he notes that with health care opportunities come significant privacy concerns. Finally, Bollier addresses how big data abuse should be handled. One idea that is proposed is to require user data to be anonymous.

Credibility. David Bollier is an author, activist, blogger and consultant in the economics, politics, and culture of the commons. Bollier is the author of 9 books on public policy issues. Between 2007 and 2010 Bollier wrote seven reports for the Aspen Institute. The aim of the Aspen Institute is to publish nonpartisan works on social issues. This work is cited by 13 other authors.

Borkar, V., Carey, M., & Li, C. (2012). Inside big data management: Ogres, onions, or parfaits? *EDBT*. Retrieved April 29, 2012 from http://www.edbt.org/Proceedings/2012-Berlin/papers/keynotes/a3-carey.pdf.

Abstract. In this paper we review the history of systems for managing "Big Data" as well as today's activities and architectures from the (perhaps biased) perspective of three "database guys" who have been watching this space for a number of years and are currently working together on "Big Data" problems. Our focus is on architectural issues, and particularly on the components and layers that have been developed recently (in open source and elsewhere) and on how they are being used (or abused) to tackle challenges posed by today's notion of "Big Data". Also covered is the approach we are taking in the ASTERIX project at UC Irvine, where we are developing our own set of answers to the questions of the "right" components and the "right" set of layers for taming the "Big Data" beast. We close by sharing our opinions on what some of the important open questions are in this area as well as our thoughts on how the data intensive computing community might best seek out answers.

Summary. The authors begin by stating that in the year 2012 businesses in diverse industries are experiencing big data. However, the authors claim that the database community has been grappling with big data since the 1980s and in particular since the 2000's when the large web properties such as Google, Amazon, and Facebook appeared.

This is consistent with Gray's fourth paradigm (Gray & Szalay, 2007). In section 2 of the paper, the authors offer a history of big data in from three perspectives: (a) big data in the database world, (b) big data in the systems world, and (c) big data today. To the authors, big data in the database world has grown over the years in response to the need to the needs of data warehouse applications. Data sizes in databases have grown from several hundred transactions per second to hundreds of thousands today. Big data in the systems world requires moving traditional databases because traditional databases (including parallel databases) are clumsy and expensive in practice. Google pioneered the Google [distributed] File System and MapReduce algorithms to handle its data loads. The open source Apache Hadoop and Hadoop Distributed File System were created to match the Google specification. In addition to the distributed file systems the MapReduce algorithms NoSQL and other tools have been developed. Today big data users choose (a) a MapReduce approach, (b) a NoSQL approach, or (c) a database sharding approach.

In the next sections, the authors consider the question of parallel database systems vs. open source systems based on Hadoop. The authors describe parallel database systems as onion-like in that they are composed of many layers of functionality with only the top-most layer exposed to the user. Also, parallel database systems are so expensive that "they have been known to make people cry" (3.1 Onions: Parallel database systems, para. 3). To the authors, Hadoop systems are like Ogres. They have some good traits but also many undesirable traits. Finally, the authors offer their solution to the Onion vs.

Ogre dilemma: a software system called ASTERIX that combines parallel processing with semi-structured data.

Credibility. Vinayek R. Borkar is a senior staff software engineer at BEA Systems, Inc. and a visiting researcher at the University of California, Irvine. Michael J. Carey is Bren Professor and vice chair of the graduate studies in Computer Science at the University of California, Irvine. Chen Li is associate professor of Computer Science at the University of California, Irvine. The EBT association is a non-profit organization for promoting research in databases and information technology. The proceedings are peer-reviewed. This work is cited by 2 authors.

Broadbent, M., & Kitzis, E. (2005). *The new CIO leader: Setting the agenda and delivering results*. Boston, MA: Havard Business School Press.

Abstract. Two converging factors--the ubiquitous presence of technology in organizations and the recent technology downturn--have brought chief information officers (CIOs) to a critical breaking point. They can seize the moment to leverage their expertise into a larger and more strategic role than ever before, or they can allow themselves to be relegated to the sideline function of "chief technology mechanic." Drawing from exclusive research conducted by Gartner, Inc., with thousands of companies and CIOs, Marianne Broadbent and Ellen Kitzis reveal exactly what CIOs must do now to solidify their credibility with the executive team and bridge the chasm that currently separates business and IT strategy.

The New CIO Leader outlines the agenda CIOs need to integrate business and IT assets in a way that moves corporate strategy forward--whether a firm is floundering, successfully competing, or leading its industry. Mandatory reading for CIOs in every firm, *The New CIO Leader* spells out how information systems can deliver results that matter--and how CIOs can become the enterprise leaders they should be.

Summary. The authors state that the role of the CIO is evolving from a role as a technologist into a role of information strategist and business leader who has the same credibility as the chief operating officer and the chief financial officer. This researcher believes that many companies recognize the need for a technology-focused officer and that is why the CTO position was created. Surprisingly, the authors state that the CIO in many businesses directly influences what path the CIO will follow, that of a technologist or a business leader. The authors then develop two competing views of the nature of the CIO's leadership: (a) supply-side leadership – concerned with determining the demand for IT services, and (b) demand-side leadership – concerned with delivering IT services that are valued by the business. In providing supply-side leadership the CIO must: (a) build an effective IT organization, (b) develop a high-caliber IT team, (c) manage risk, and (d) communicate about IT performance. In providing demand-side leadership the CIO must: (a) understand the fundamentals of the business's environment, (b) create an IT vision, (c) manage expectations of IT within the business, (d) establish effective IT governance, and (e) provide an integrated business-IT strategy. The authors conclude that a CIO's evolution is not an easy one. But they state that if a CIO fulfills all elements of supply-side and demand-side leadership they will likely become key business executives.

Credibility. Marianne Broadbent is an execute advisor, speaker, author and management consultant in leadership issues. Prior to that Broadbent was senior executive vice president of new product development at Gartner, Inc. Broadbent has also served as associate dean at the Melbourne Business School, University of Melbourne. Dr. Ellen S. Kitzis is group vice president of Executive Programs at Gartner, Inc. Harvard Business School Press publishes well-regarded and widely read authoritative books on management issues. This work is cited by 108 authors.

Friendly, M. (2009). *Milestones in the history of thematic cartography, statistical graphics,* ***and*** *data visualization*. Retrieved June 11, 2012 from http://euclid.psych.yorku.ca/SCS/Gallery/milestone/milestone.pdf.

Abstract. The graphic portrayal of quantitative information has deep roots. These roots reach into histories of thematic cartography, statistical graphics, and data visualization, which are intertwined with each other. They also connect with the rise of statistical thinking up through the 19th century, and developments in technology into the 20th century. From above ground, we can see the current fruit; we must look below to see its pedigree and germination. There certainly *have* been many new things in the world of visualization; but unless you know its history, everything might seem novel.

Summary. Friendly presents the history of data visualization in a format that can be analyzed using Gray's four scientific paradigms (Gray 2007). During Gray's first paradigm (a thousand years ago) Friendly identifies progress in map making, primitive bar graphs, and tabular representations. During Gray's second paradigm (a few hundred years ago) Friendly identifies the use of visual representations of phenomenon, the usage of statistical data and complex graphing. Friendly refers to Gray's third paradigm (last few decades) as the "Modern Dark Ages" (Modern Dark Ages, para. 1) where computational simulation and statistical models replaced visualization techniques.

Finally in Gray's fourth paradigm (today) Friendly reports that visualization is an integral part of big data and includes the preprocessing of data and interactive graphical display**.**

Credibility. Michael Friendly is professor of Psychology, chair of the graduate program in Quantitative Methods at York University, and an associate coordinator with the Statistical Consulting Service at York University. Friendly holds a doctorate in Psychology from Princeton University and is the author of three books on data visualization and numerous online papers. The Statistical Consulting Service is hosting this work for download. This work is cited by 77 authors.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery* [Kindle version]. Redmond, WA: Microsoft Research.

Abstract. This book presents the first broad look at the rapidly emerging field of data-intensive science, with the goal of influencing the worldwide scientific and computing research communities and inspiring the next generation of scientists. Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets. The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud-computing technologies. This collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Summary. The book starts by voicing Jim Gray's warning that big data costs are ¼ to ½ composed of software costs. Book sections are devoted to (a) earth and environment (b) health and well-being, (c) scientific infrastructure and (d) scholarly communication. In the earth and environment section, key ideas include that (a) big data must process sensor data in real time, (b) models do not account for environmental anomalies and therefore are not very useful, and (c) climate change data needs to be analyzed using statistics and correlation in order to achieve rapid scientific understanding. In the health and wellbeing section key ideas are that (a) medical discovery will translate into medical practice, (b) expert systems will target improved world health, (c) a complete circuit diagram of the brain will be produced, (d) smart medical instruments such as computational microscopes will be created, and (e) "health avatars" (A unified modeling approach to data-intensive healthcare, para. 9) formed by the aggregation of an individual's health records will improve an individual's health. In the scientific infrastructure section key ideas include (a) the intersection of information technology and science, (b) next generation genomic sequencing, (c) advances in parallel programming and cloud computing, and (d) automated workflow tools for acquiring data and advanced visualization. In the scholarly communication section key ideas include (a) the existence of vast reference data sets, (b) data that is subject to ongoing computation and refinement, (c) semantic text enhancement, and (d) machine readable scholarly article databases.

Credibility. Tony Hey is vice president of Microsoft Research Connections, a division of Microsoft Research. Prior to his current position Hey was director of the UK's e-Science Initiative. He holds a doctorate in Physics from Oxford. Stewart Tansley is a Director of Microsoft Research Connections, a division of Microsoft Research. Tansley holds a doctorate in Artificial Intelligence from Loughborough University, UK. Kristen M. Tolle is a Director of Microsoft Research Connections, a division of Microsoft

Research. Tolle holds a doctorate in Management of Information Systems from the University of Arizona. The stated goal of Microsoft Research is the advance the state of the art in general computing and not just for Microsoft corporate purposes. This book as well as many other publications is available for free download. This book is cited by 270 other authors.

Linoff, G., & Berry, M. (2011). *Data mining techniques: For marketing, sales, and customer relationship management*. Indianapolis, IN: Wiley Publishing.

Abstract. When Berry and Linoff wrote the first edition of "Data Mining Techniques" in the late 1990s, data mining was just starting to move out of the lab and into the office and has since grown to become an indispensable tool of modern business. This new edition-- more than 50% new and revised-- is a significant update from the previous one, and shows you how to harness the newest data mining methods and techniques to solve common business problems. The duo of unparalleled authors share invaluable advice for improving response rates to direct marketing campaigns, identifying new customer segments, and estimating credit risk. In addition, they cover more advanced topics such as preparing data for analysis and creating the necessary infrastructure for data mining at your company. Features significant updates since the previous edition and updates you on best practices for using data mining methods and techniques for solving common business problems. Covers a new data mining technique in every chapter along with clear, concise explanations on how to apply each technique immediately. Touches on core data mining techniques, including decision trees, neural networks, collaborative filtering, association rules, link analysis, survival analysis, and more. Provides best practices for performing data mining using simple tools such as Excel.

Summary. The author states that small businesses build customer relationships naturally via one to one interactions over a long period of time. Large businesses with millions of customers build these lasting relationships with their customers over time with data mining. Most of the chapters of the book are devoted to detailed data mining techniques. The author does include higher-level chapters on (a) the data mining process, (b) a tutorial on data, (c) data warehousing, and (c) social network data mining. The author also states that today businesses are motivated to mine data because (a) data is being produced at a rapid rate, (b) data is being stored a manner that lends itself to data mining, (c) computing is inexpensive, and (d) businesses are seeking competitive advantage in employing customer relationship management.

Credibility. Gordon S. Linoff and Michael J. A. Berry are founders of Data Miners, Inc. and are recognized experts in the field of data mining. They state that they maintain a vendor-neutral stance in their practice and in their writings. John Wiley is a publisher of authoritative books on a variety of subjects. This work is cited by 2602 authors.

Lutchen, M. (2004). *Managing IT as a business: A survival guide for CEOs*. Hoboken, NJ: John Wiley and Sons.

Abstract. Information technology ranks in the top five of most business's expenditures. Yet information technology is the least understood and poorly managed areas in business. While most executives realize the importance of information technology, few understand how to leverage IT strategically and how to use it as driver of business success. *Managing IT as a business* provides executives with practical advice on how to unleash the full potential of this critical function so that companies can derive maximum benefits (Lutchen, front bookcover flap).

Summary. The author begins by stating that IT historically has been given large budgets yet it has lagged in focus and strategic vision in the business. In order to fix this gap in performance companies have either attempted to fix IT with a strategic or a technological focus. The author lays out an alternative plan based on the following six dimensions: (a) alignment to business goals, (b) resiliency of resources from disruption, (c) integration of emerging technology, (d) support of IT (staffing, finance, communication), (e) operations, and (f) leveraging IT assets across the business. The author then analyzes and presents methods that a business can employ to bring IT into alignment with the business as the following: (a) make the CIO a member of the executive board, (b) link IT strategy to corporate strategy, (c) operate IT as a business profit center, (d) understand the true costs of IT, (e) manage outcomes by establishing and tracking performance metrics, (f) fund IT as an annual expense instead of with capital projects, (g) adopt standardized systems, and (h) communicate about IT to the business. The author concludes this approach works to bring IT into alignment with business. However, the nature of business is dynamic and IT must be part of that dynamism. It is the opinion of this researcher that three areas that Lutchen examines explain directly how the CIO and CTO can extract value from big data to the business: (a) alignment to business goals (b) integration of emerging technology, and (c) leveraging IT assets across the business.

Credibility. Mark D. Lutchen currently leads the IT Business Risk Management Practice at PricewaterhouseCoopers. Lutchen is also a speaker and consultant with 30 years of experience. Lutchen is a graduate of the Harvard, INSEAD and MIT/Sloan School executive development programs. PricewaterhouseCoopers is a business consulting company and underwrites the authoring of business books as part of its mission. This work is cited by 70 authors.

Warden, P. (2011). *Big data glossary* [Kindle version]. Sebastopol, CA : O,Reilly Media.

Abstract. There has been a massive amount of innovation in data tools over the last few years, thanks to a few key trends: Learning from the Web Techniques originally developed by website developers coping with scaling issues are increasingly being applied to other domains. Google has proven that research techniques from computer science can be effective at solving problems and creating value in many real-world situations. That's led to increased interest in cross-pollination and investment in academic research from commercial organizations. Now that machines with a decent amount of processing power can be hired for just a few cents an hour, many more people can afford to do large-scale data processing. They can't afford the traditional high prices of professional data software, though, so they've turned to open source alternatives. These trends have led to a Cambrian explosion of new tools, which means that when you're planning a new data project, you have a lot to choose from. This guide aims to help you make those choices by describing each tool.

Summary. The author states that the explosion of big data has spawned a lexicon of terminology that rivals the size of big data itself. The author has captured the essence of big data technology into a specialized Encyclopædia with sections on (a) NoSQL databases, (b) MapReduce, (c) storage and servers, (e) data processing and machine learning, (g) analytics, and (h) data acquisition and serialization. It is the opinion of this researcher that understanding big data project costs is predicated upon being familiar with big data technology terms and capabilities.

Credibility. Pete Warden is the founder of the open-source project OpenHeatMap. Warden also writes on big data and visualization. O'Reilly publishes peer-reviewed books on computer technology.

Data-Intensive Computing and Relationships to Big Data

Data intensive computing is composed of common components of (a) data acquisition, (b) data management, (c) modeling and simulation, (d) algorithms, (e) information analytics, and (f) computing platforms (Kouzes et al., 2009). Big data is a type of data intensive computing that requires new common components to meet enormous data challenges (Gorton et al., 2008, April).

Ali, A., Chen, Z., & Lee, J. (2008). Web-enabled platform for distributed and dynamic decision-making systems. *International Journal of Advanced Manufacturing Technology, 38*, 11-12. Retrieved June 1, 2012 from http://www.springerlink.com/content/n2333763960716q6/.

Abstract. With the advent of internet and wireless technologies, real-time remote monitoring and control is becoming an essential need for meeting highly dynamic business objectives. At the same time, web-enabled platforms are required to perform remote monitoring with efficiently and effectively. Recent progress on e-manufacturing applications addresses the needs for better integration between factory floor and enterprise systems. This paper presents a web-enabled platform which focuses on web-enabled intelligence to enable products and systems to achieve near-zero-downtime performance through device-to-business (D2B) platform. The proposed web-enabled platform can effectively minimize the massive information bottleneck that exists between plant floor and information systems. Case studies are presented to determine how effectively web-enabled industrial system can be used in factory floor as well as business decision-making. Manufacturers and users will benefit from the increased equipment and process performance with the effective implementation of the developed web-enable platform.

Summary. A need exists in manufacturing to bridge the gap between the business system and the manufacturing systems. This need is best met through a web-enabled system that combines real time data with flexible machining and assembly cells to create an agile data-driven manufacturing environment. Big data is harvested in the e-manufacturing system by expert systems to optimize the production schedule and to prevent defects. Data manipulation and processing occur in a distributed fashion to efficiently use the data without the need to re-transmit data. Synchronization between the e-manufacturing system and the business system handles key factory floor transactions such as requests for parts, service and replacement tools. While the authors present a novel approach to using big data on the factory floor, this researcher believes that their analysis, however, is biased towards the use of traditional databases. An updated version of this work would possibly recommend the use of NoSQL databases.

Credibility. Ahad Ali is a professor of Industrial Engineering at the University of Puerto Rico – Mayaguez. Ziafeng Chen is with Automated Precision, Inc. of Rockville, Maryland. Jay Lee is Ohio Eminent Scholar and L.W. Scott Alter Chair professor at the University of Cincinnati and is founding director of the National Science Foundation Industry/University Cooperative Research Center on Intelligent Maintenance Systems.

The International Journal of Advanced Manufacturing Technology publishes peer-reviewed papers. This work is cited by 4 other authors.

Bell, G. (2009). Foreword. In T. Hey, S. Tansley & K. Tolle. *The fourth paradigm: Data intensive scientific discovery*. Redmond, WA: Microsoft Research. Retrieved June 11, 2012 from http://research.microsoft.com/en-us/collaboration/fourthparadigm/contents.aspx.

Abstract. There is a sea change happening in academic research -- a transformation caused by a data deluge that is affecting all disciplines. Modern science increasingly relies on integrated information technologies and computation to collect, process, and analyze complex data. It was Ken Wilson, Nobel Prize winner in physics, who first coined the phrase "Third Paradigm" to refer to computational science and the need for computational researchers to know about algorithms, numerical methods, and parallel architectures. However, the skills needed for manipulating, visualizing, managing, and, finally, conserving and archiving scientific data are very different. "The Fourth Paradigm" is all about the data and the computational systems needed to manipulate, visualize, and manage large amounts of scientific data. A wide variety of scientists — biologists, chemists, physicists, astronomers, engineers – require tools, technologies and platforms that seamlessly integrate into standard scientific methodologies and processes. This talk will illustrate the far-reaching changes that this new paradigm will have on scientific discovery and the role that the Cloud, semantic computing and research repositories will have in this new landscape.

Summary. The forward begins with the analogy that data-intensive computing in the fourth paradigm is what the printing press was to printing, i.e., data-intensive computing is in its infancy and will likely take decades to mature. Science lags behind the commercial sector on data-intensive computing mainly because it's easier to create usable commercial data-intensive computing systems. Certain new developments have arisen in data-intensive computing and include (a) scientific electronic data that was previously very fragile and subject to loss will soon live on ubiquitously and are freely available, and (b) automated sensors that produce data faster than it can be consumed. Additional funding needs to be provided to curate the data otherwise data will be forever lost. Public digital libraries are emerging that are central repositories for experimental data.

Credibility. Gordon Bell is a principle researcher at Microsoft Research. Prior to that Bell was vice president of Research and Development at Digital Equipment Corporation. Bell is the author of numerous books and articles and is a fellow of the American Academy of Arts and Sciences. The stated goal of Microsoft Research is to advance the state of the art in general computing and not just free for Microsoft corporate purposes. This book as well as many other publications is available for free download. This book is cited by 270 other authors.

Bell, G., Hey, T., & Szalay, A. (2009, March). Beyond the data deluge. *Science,* 1297-1298. Retrieved June 11, 2012 from http://www.sciencemag.org/content/323/5919/1297.summary.

Abstract. Since at least Newton's laws of motion in the 17th century, scientists have recognized experimental and theoretical science as the basic research paradigms for understanding nature. In recent decades, computer simulations have become an essential third paradigm: a standard tool for scientists to explore domains that are inaccessible to theory and experiment, such as the evolution of the universe, car passenger crash testing, and predicting climate change. As simulations and experiments yield ever more data, a fourth paradigm is emerging, consisting of the techniques and technologies needed to perform data-intensive science. For example, new types of computer clusters are emerging that are optimized for data movement and analysis rather than computing, while in astronomy and other sciences, integrated data systems allow data analysis and storage on site instead of requiring download of large amounts of data.

Summary. This article explains the particular complexities of data-intensive computing in scientific applications. It also sounds a warning that data-intensive scientific computing is nearing the breaking point due to storage constraints, analytic tools, and trained professionals. The authors state that Gray's fourth paradigm is an emerging paradigm in data-intensive computing. They further state that data-generating capabilities of instruments, sensors, and advanced computers are outpacing the ability to store the generated data. Gray recognized this problem early on and proposed building specialized storage appliances to house the data. Newer architectures have been proposed as well. Data-intensive computing now embraces large public networks and commodity hardware. Yet data-intensive scientific computing lags behind commercial data-intensive computing primarily from the lack of specialists and data analysis tools.

Credibility. Gordon Bell is a principle researcher at Microsoft Research. Prior to that, Bell was vice president of Research and Development at Digital Equipment Corporation. Bell is the author of numerous books and articles and is a fellow of the American

Academy of Arts and Sciences. Tony Hey is vice president of Microsoft Research Connections, a division of Microsoft Research. Prior to his current position, Hey was director of the UK's e-Science Initiative. He holds a doctorate in Physics from Oxford. Stewart Tansley is a Director of Microsoft Research Connections, a division of Microsoft Research. Alex Szalay is Alumni Centennial professor of Physics and Astronomy at the Johns Hopkins University. *Science* publishes peer-reviewed journals. This article is cited by 5 authors.

Cohen J., Dolan, B., Dunlap, M. , Hellerstein, J. M., & Welton, C. (2009). MAD skills: New analysis practices for big data. *Proceedings of the VLDB Endowment, 2*(2), 1481-1492. Retrieved June 1, 2012 from http://dl.acm.org/citation.cfm?id=1687553.1687576.

Abstract. As massive data acquisition and storage becomes increasingly affordable, a wide variety of enterprises are employing statisticians to engage in sophisticated data analysis. In this paper we highlight the emerging practice of Magnetic, Agile, Deep (MAD) data analysis as a radical departure from traditional Enterprise Data Warehouses and Business Intelligence. We present our design philosophy, techniques and experience providing MAD analytics for one of the world's largest advertising networks at Fox Audience Network, using the Greenplum parallel database system. We describe database design methodologies that support the agile working style of analysts in these settings. We present dataparallel algorithms for sophisticated statistical techniques, with a focus on *density* methods. Finally, we reflect on database system features that enable agile design and flexible algorithm development using both SQL and MapReduce interfaces over a variety of storage mechanisms.

Summary. As data-intensive computing has become more prevalent, the author writes that the analysis of big data is migrating beyond the traditional data warehouse and business intelligence. The author proposes a new analysis technique called *MAD* – Magnetic, Agile and Deep – as a departure from traditional analysis systems. In the introduction, the author explains that business data warehouses are designed in a very disciplined and rigorous fashion, which results in a very reliable but completely inflexible system. The relatively modest cost of big data today is allowing groups within a business to perform data-intensive computing outside of the business warehouse. The authors further elaborate on MAD as improving business intelligence by (a) *magnetically* pulling in data even from unapproved sources, (b) *agilely* using data as it becomes available and (c) *deeply* using data beyond what is capable in the traditional data warehouse. In order to meet the requirements of MAD new approaches to data systems need to be taken such as:

- During the project requirements phase a more iterative approach to data analysis is used instead of the traditional architected approach.
- The emphasis is on getting data into the big data system as early as possible.
- A big data development environment is set up so that new ideas and approaches can be introduced into production rapidly.
- Statistics processing should use parallel processing to be effective.
- Data input into big data systems needs to done efficiently.
- Allow data scientists to use analysis tools that they are comfortable with whether they are programming languages like Java or statistical analysis packages like SAS.

Credibility. Jeffrey Cohen and Caleb Welton are employed by Greenplum, a division of EMC, a company that specializes in big data analytic systems. Brian Dolan is employed by Fox Audience Network, a user of Greenplum. Mark Dunlap is employed by Evergreen Technologies, an IT consulting company. Joseph M. Hellerstein is Chancellor's professor of Electrical Engineering and Computer Science at the University of California, Berkeley. He has a doctorate from the University of Wisconsin and is an Association of Computing Machines fellow. Although this article discusses Greenplum's products and presents an application of Greenplum at Fox Audience Network, the article reflects aspects of big data analytics that appear to be free from bias. This article is cited by 47 authors.

Kouzes, R., Anderson, G., Elbert, S., Gorton, I., & Gracio, D. (2009). The changing paradigm of data-intensive computing. *Computer*, *42*(1), 26-34. doi: 10.1109/MC.2009.26.

Abstract. Through the development of new classes of software, algorithms, and hardware, data-intensive applications provide timely and meaningful analytical results in response to exponentially growing data complexity and associated analysis requirements.

Summary. The authors begin by stating that all existing definitions of data-intensive computing focus on the handling of massive amounts of data. However, due to new requirements and applications, a broader definition of data-intensive computing is required in order to fully explain how data-intensive computing is evolving. The authors then present some new data-intensive computing applications such as the North American power grid operations and Facebook where vast amounts of data processing are present. The authors present several data-intensive computing styles that have emerged recently: (a) the Data-processing Pipeline compresses from a raw state to a final state ready for analytics, (b) the Data Warehouse creates a highly structured aggregation of data ready for analytics, and (c) the Data Center uses MapReduce or Hadoop to create distributed and flexible big data. As data-intensive computing has evolved into big data many challenges exists such (a) how to route and process incoming data streams, (b) data integrity, (c) parallel database scaling, (d) metadata management, (e) big data integration, (f) the resources required to analyze the data including machine learning. In order to successfully implement big data an integrated design needs to combine mathematical, statistical and computer science capabilities. Key factors that will determine the future direction of big data are (a) advances in computer hardware (b) digital sensor data flow rates, (c) and MapReduce (as the open source Hadoop) adoption.

Credibility. Richard T. Kouzes is a laboratory fellow at Pacific National Laboratory. He received a PhD in physics from Princeton University and is a Fellow of the IEEE and the American Association for the Advancement of Science. Gordon A. Anderson is an associate director for scientific resources at Pacific Northwest National Laboratory. He received a BS in engineering from Washington State University. Stephen T. Elbert is a manager at Pacific Northwest National Laboratory. He received a PhD in computational chemistry from the University of Washington. Ian Gorton is an associate division director at Pacific Northwest National Laboratory. Deborah K. Gracio is a computational and statistical analytics division director at Pacific Northwest National Laboratory. The IEEE Computer Society publishes peer-reviewed articles. This article is cited by 23 authors.

Current Opportunities of Big Data System Applications

Big data will be widely adopted in the commercial, science, medical and health, and government sectors (Bryant et al., 2008). Manyika et al. (2011) write that certain economic sectors such computer, electronic products, finance, insurance and government will see the most adoption of big data but eventually all sectors will have to adopt big data.

Bryant, R. E., Katz, R. H., & Lazowska, E. D. (2008). *Big-data computing: Creating revolutionary breakthroughs in commerce, science and society*. Computing Research Association. Retrieved June 11, 2012 from http://www.cra.org/ccc/docs/init/Big_Data.pdf.

Abstract. Advances in digital sensors, communications, computation, and storage have created huge collections of data, capturing information of value to business, science, government, and society. For example, search engine companies such as Google, Yahoo!, and Microsoft have created an entirely new business by capturing the information freely available on the World Wide Web and providing it to people in useful ways. These companies collect trillions of bytes of data every day and continually add new services such as satellite images, driving directions, and image retrieval. The societal benefits of these services are immeasurable, having transformed how people find and make use of information on a daily basis.

Summary. The authors state that advances in sensors, communications, computation, and storage have created big data. Early on, big data was utilized primarily by search engines, but now, businesses such as Walmart mine their big data being fed by data streams from cash register and warehouse sensors. Big data is possible by advances in:

(a) sensor networks, (b) computer networks, (c) data storage, (d) distributed computing, (e) cloud computing, and (f) data analysis algorithms. Many aspects of big data are advancing normally (i.e. they are not problematic); however, several aspects of big data require more focused attention including: (a) high-speed networking, (b) more powerful distributed computing, (c) extending the reach of cloud computing, (d) machine learning, (e) pervasive availability of big data computing, and (f) security and privacy. The authors report that leadership in big data is lagging in university research and governmental agencies. Finally, the authors state near-term and long-term benefits of big data. In the near term the authors recommend the following: (a) do not overspend on computer hardware which will quickly become obsolete, (b) invest in high capacity networking equipment that has a longer lifetime, (c) invest in public big data networking at universities and governmental agencies, and (d) increase the National Science Foundation research short term budget for big data. In the long term the authors recommend the following: (a) increase the National Science Foundation research budgets to fully fund new programs, (b) provide funding to construct several massive data centers for government projects, (c) increase DARPA funding, (d) provide funding to secure big data sites against cyber-attack, (e) turn governmental agencies into consumers of big data, and (f) increase funding for better computer networks.

Credibility. Randal E. Bryant is University professor of Computer Science at Carnegie Mellon University. He holds a doctorate in Applied Mathematics from MIT. Randy H. Katz is a United Microelectronics Corporation Distinguished professor of Electrical Engineering and Computer Science at the University of California, Berkeley. He holds a doctorate in Applied Mathematics from the University of California, Berkeley. Edward

D. Lazowska holds the Bill & Melinda Gates Chair in Computer Science & Engineering at the University of Washington. He holds a doctorate from the University of Toronto. The Computer Research Association aims to promote great understanding of computing and to influence computing policy. One of their missions is to publish peer-reviewed papers. This article is cited by 6 authors.

Eagle, N. (2010). Big data, global development, and complex social systems. *Proceedings of the eighteenth ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE '10),* 3-4. Retrieved June 2, 2012 from http://www.legacyforhealth.org/PDF/Eagle.pdf.

Abstract. Petabytes of data about human movements, transactions, and communication patterns are continuously being generated by everyday technologies such as mobile phones and credit cards. In collaboration with the mobile phone, internet, and credit card industries, Eagle and colleagues are aggregating and analyzing behavioral data from over 250 million people from North and South America, Europe, Asia and Africa. Eagle discusses projects arising from these collaborations that involve inferring behavioral dynamics on a broad spectrum of scales from risky behavior in a group of MIT freshman to population-level behavioral signatures, including cholera outbreaks in Rwanda and wealth in the UK. The research group is developing a range of large-scale network analysis and machine learning algorithms that will provide deeper insight into human behavior.

Summary. The author begins by stating that researchers are good at analyzing small, static models but not so good at analyzing large, continuously changing models that have varied outcomes. The author advises moving away from theory and more towards an outcomes-based approach based on understanding network topology. Big data is being collected so that extremely small populations can be examined to see how they affect a larger population. For example, a 50 household population can be examined to see how Respiratory Syncytial Virus spreads. The author even proposes that people could be studied as particles to provide new laws governing human behavior. Finally, the author asks the question: Is big data a new social science?

Credibility. Nathan Eagle is an Omidyar Fellow at the Santa Fe Institute and a visiting assistant professor in the MIT Media Lab. Eagle holds a doctorate in Media Arts and Sciences from the MIT Media Lab. The abstract for this conference is available from the ACM for citing.

Franks, B. (2012). *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics* [Kindle edition]. Hoboken, NJ: John Wiley & Sons, Inc.

Abstract. To stay ahead of the pack, you need to tackle big data today. Discover everything you need to improve productivity, create value, stay competitive, spot new business trends, and generate exciting analytics with *Taming the big data tidal wave*.

Summary. The author states that big data is becoming a hallmark of the start of the 21st century where big data is being consumed and utilized by more and more businesses. The author provides some important observations about big data: (a) the existence of a big data set does not provide any credibility to the data, (b) big data is typically machine generated and represents a new source of data, (c) big data can be ugly and messy data that requires extreme care when analyzing, (d) big data is risky in terms of privacy, (e) big data needs to be tamed to make it useful, and (f) big data is powerful when mixed with traditional data. The author investigates the rise of big data and how to tame big data with technology and people. Once businesses tame big data it leads to a condition the author describes as an analytics culture in the business. The author concludes that big data is a permanent addition to the business culture and that business leaders need to recognize this and take action to bring big data into the business.

Credibility. Bill Franks is Chief Analytics Officer at Teradata, Inc., a provider of data warehouse and analytics solutions. Franks is also a faculty member of the International Institute for Analytics that is dedicated to the advancement of analytics in business. Franks is employed by a provider of analytics solutions, therefore the possibility of bias exists. However, Frank's basic premise – that big data and analytics are critically important to business - is supported by many researchers and experts in big data. John Wiley is a publisher of authoritative books on a variety of subjects.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. (2010). Big data, analytics, and the path from insights to value. *MIT Sloan Management Review*, *62*(2). Retrieved June 2, 2012 from http://sloanreview.mit.edu/the-magazine/2011-winter/.

Abstract. How the smartest organizations are embedding analytics to transform information into insight and then action. Findings and recommendations from the first annual New Intelligent Enterprise Global Executive study.

Summary. The authors conducted a survey of 3000 business executives to see how well they use big data analytics. Their finding is that top business performers use analytics five times more frequently the lowest business performers. Other findings of their survey are: (a) improvement in analytics is a top business priority, (b) six out of ten survey respondents believe that competitive differentiation in business by adopting innovation is a top business challenge, and (c) six out of ten survey respondents believe that they are collecting more data than they can consume. The authors make some observations: (a) top business performers are saying that business analytics is an important business differentiator, (b) businesses are at different levels of maturity regarding big data, (c) a lack of understanding of the data and not the data itself is the biggest impediment to using big data, (d) data must be easier to understand and to use, and (e) business leaders must adopt newer methodologies to successfully use big data. Finally, the authors offer these recommendations:

- Focus on the biggest and highest-value opportunities because they command the best resources.
- Start with questions and not data in the design of big data.
- Embed big data analytics into business processes.
- Keep existing capabilities while adding new ones.
- Build data sub systems without losing sight of the overall design goals.

Credibility. Steve LaValle is the global strategy leader for IBM's Business Analytics and Optimization service line. Eric Lesser is the research director and North American leader of the IBM Institute for Business Value. Rebecca Shockley is the business analytics and optimization global lead for the IBM Institute for Business Value. Michael S. Hopkins is editor-in-chief of MIT Sloan Management Review. Nina Kruschwitz is an editor and the special projects manager at MIT Sloan Management Review. MIT Sloan Management Review is a well-regarded business periodical. This work is cited by 10 authors.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity.* McKinsey Global Institute. Retrieved June 11, 2012 from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

Abstract. The amount of data in our world has been exploding, and analyzing large data sets—so-called big data—will become a key basis of competition, underpinning new waves of productivity growth, innovation, and consumer surplus, according to research by MGI and McKinsey's Business Technology Office. Leaders in every sector will have to grapple with the implications of big data, not just a few data-oriented managers. The increasing volume and detail of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future.

Summary. The authors state that big data is becoming increasingly prevalent due to the proliferation of data sources (smart sensors, phones, etc.) and software applications (social media, mobile, web sites). The authors claim that while there are many industry sectors that are using big data in their business activities, many are not. Also, a shortage of specialists in big data exists, and that is hampering the adoption of big data. Key ideas from the book’s sections are the following:

Mapping global data: Growth and value creation. Data is growing at an exponential rate, but the data growth in each industry sector is not uniform. IT has contributed greatly to productivity growth, and big data is the phase of productivity growth.

Big data techniques and technologies. Techniques such as data mining and machine learning derive from the statistics and computer science field but expand on these extant techniques to deal with big data. Also many new technologies are now available that work well with big data such as cloud computing, distributed systems, and Hadoop.

The transformative potential of big data in five domains. The five domains chosen by the authors represent 40% of global GDP in 2010. In health care costs are rising faster that economic growth, and targeted uses of big data can be used to help contain health care costs. In the public sector governments are under extraordinary pressure to increase productivity, and big data can be used to direct public resources to areas that do the most good under tightening resource (tax receipt) constraints. Retailers face tight profit margins, and big data can be used to improve margins by using big data to serve customer’s needs. Manufacturing struggles to contain costs in a globally competitive economy, and big data can be used to tighten the integration of supplier networks in order to be more responsive to change. Personal location data a nascent industry but the authors believe that the industry will grow to as much as $700 billion in worldwide annual revenues.

The authors conclude by presenting key findings: (a) big data creates real value to business, (b) some business sectors are experiencing gains from big data faster than others, (c) some geographic areas (particularly North America and Europe) are achieving more rapid growth in big data than others, (d) there is a growing shortage in big data talent, and (e) big data has a growing number of issues (such as privacy) that need to be addressed. The authors offer advice to business leaders: (a) know your big data inventory, (b) identify opportunities and threats in using big data, (c) build internal big data capabilities, (d) develop IT strategy towards big data, and (e) address big data security and privacy issues. Finally the authors offer advice to public policy makers: (a) increase funding for big data education, (b) increase research in targeted big data areas, and (c) continue investing in information and communication infrastructure.

Credibility. The authors are employees of McKinsey Global Institute, a research unit of the McKinsey management consulting firm. All research from McKinsey Global Institute is funded by McKinsey alone; no work is commissioned from business, government or other entities. This work is cited by 22 authors.

Ratner, B. (2011). *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data, second edition.* CRC Press: New York.

Abstract. The second edition of a bestseller, *Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data* is still the only book, to date, to distinguish between statistical data mining and machine-learning data mining. The first edition, titled *Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data*, contained 17 chapters of innovative and practical statistical data mining techniques. In this second edition, renamed to reflect the increased coverage of machine-learning data mining techniques, the author has completely revised, reorganized, and repositioned the original chapters and produced 14 new chapters of creative and useful machine-learning data mining techniques. In sum, the 31 chapters of simple yet insightful quantitative techniques make this book unique in the field of data mining literature.

Summary. The author states that statisticians have been using data mining for decades through the use of exploratory data analysis in analyzing big data sets. Today the author believes that data mining is a combination of statistical analysis, big data and machine learning. The author gradually develops themes in data mining from very simple ones such as correlation to relatively complex topics in machine learning and the authors' own GenIQ machine learning model. In the four chapters of the book the author presents applications of the GenIQ model to real-world big data problems.

Credibility. Bruce Ratner is founder and president of DM STAT-1 CONSULTING, which provides data mining and statistical services to a variety of business sectors such as health care, finance and retail. Ratner holds a doctorate in applied mathematics. CRC press is a publisher of peer-reviewed technical and scientific books. The first edition of this work is cited by 33 authors.

Schreyogg, G., & Kliesch-Eberl, M. (2007). How dynamic can organizational capabilities be? Towards a dual-process model of capability dynamization. *Strategic Management Journal*, *28*(9), 913-933. doi: 10.1002/smj.613

Abstract. The recent discussion in the field of strategic management broadly favors the idea of dynamic capabilities in order to overcome potential rigidities of organizational capability building. The major question addressed in this paper is whether capabilities can actually be conceived as being in flux—and if so, to what extent and in which way? After briefly recapitulating the distinguishing features of organizational capabilities, path dependency, structural inertia, and commitment are identified as the main capability-rigidity drivers causing a managerial dilemma. In the search for a resolution of this dilemma different approaches of dynamic capabilities are identified and discussed. The analysis shows that the approaches suffer from inherent conceptual contradictions: the dynamization runs the risk of dissolving the original idea and strength of organizational capability building. Ultimately, capabilities would lose the strategic power attributed to them in the resource-based view. The last section of this paper therefore aims to develop an alternative approach, which aims at preserving the original merits of organizational capability and solving the rigidity issue not by integrating a dynamic dimension into the capability construct but rather by establishing a separate function ('capability monitoring').

The suggestions mount up to a tier solution. Its logic builds on the dynamics of countervailing processes and second-level observation.

Summary. The authors report that the concept of organizational capability has undergone many revisions and to some extent is identified by the type of economic model the organization decides to use. The authors report that the term *organizational capability* has many pseudonyms such as core competency and best practices. The authors believe three key characteristics about organizational capability are important: (a) problem solving and decision making under uncertainty, (b) practicing decision outcomes and ingraining successful methods in the business, and (c) building and sustaining reliable capabilities over time. However, the authors report that three key impediments that businesses must overcome to implement organizational capability changes are (a) structural inertia, (b) historical decisions and practices, and (c) commitment to implement change. A paradox exists when organizations adopt new capabilities in that on the one hand organizational capabilities enable businesses to efficiently manage their resources and on the other hand allow businesses take on new risks in choosing a capability plan over alternative plans. The authors propose that the answer to this paradox is to build a "dual-process model" (p. 925) where a second business capability is put into place to monitor the primary capability and to send alerts when the primary capability is not meeting goals.

Credibility. Georg Schreyogg is professor of Business and Economics and chair of Organization and Leadership at the Freie Universitat in Berlin, Germany. Schreyogg holds a doctorate from Universitat Erlangen-Nurnberg. Martina Kliesch-Eberl is a researcher at the Freie Universitat in Berlin, Germany. The Strategic Management journal publishes peer-reviewed articles. This work is cited by 272 authors.

Witten, I., & Elbe, F., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Abstract. Data Mining: Practical Machine Learning Tools and Techniques offers a thorough grounding in machine learning concepts as well as practical advice on applying machine learning tools and techniques in real-world data mining situations. This highly anticipated third edition of the most acclaimed work on data mining and machine learning will teach you everything you need to know about preparing inputs, interpreting outputs, evaluating results, and the algorithmic methods at the heart of successful data mining. Thorough updates reflect the technical changes and modernizations that have taken place in the field since the last edition, including new material on Data Transformations, Ensemble Learning, Massive Data Sets, Multi-instance Learning, plus a new version of the popular Weka machine learning software developed by the authors.

Witten, Frank, and Hall include both tried-and-true techniques of today as well as methods at the leading edge of contemporary research.

Summary. The authors begin by stating that data mining is about analyzing data already present in a database to discover patterns that are present in the data. The authors then state that the key elements to data mining are identifying and processing the structural data patterns – those patterns that contain information on which to base decisions. Often the data mining is performed with machine learning. In part I of the book, the authors introduce data mining process as follows: (a) start with data inputs, (b) apply data mining algorithms, and (c) produce outputs such statistical results, analytics, and rules. In part II of the book, the authors present advanced data mining techniques that have application in real world problems. In part III of the book, the authors present the Weka open source data mining software.

Credibility. Ian Witten is professor of Computer Science at the University of Waikato, New Zealand. Witten has a doctorate in Electrical Engineering from the University of Sussex, England. Eibe Frank is also a professor of Computer Science at the University of Waikato, New Zealand. Mark Hall is Honorary Research Associate in the department of Computer Science at the University of Waikato, New Zealand. Hall has a doctorate in Computer Science from the University of Waikato, New Zealand. Morgan-Kaufman is a publisher of peer-reviewed textbooks. This work is cited by 16297 authors.

Current Limitations of Big Data System Applications

Lynch (2008, September) proposes that big data is constrained by (a) cost, (b) lack of standards for data description and exchange, (c) data preservation. Bantleman (2012, April 16) states that a shortage of data science professionals is constraining the adoption of big data. Jacobs (2009) states that big data reaches limitations imposed by computer hardware, and the implementation of big data requires costly distributed computing.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society,* 662-679. doi: 10.1080/1369118X.2012.678878.

Abstract. The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing genetic sequences, social media interactions, health records, phone logs, government records, and other digital traces left by people. Significant questions emerge. Will large-scale search data help us create better tools, services, and public goods? Or will it usher in a new wave of privacy incursions and invasive marketing? Will data analytics help us understand online communities and political movements? Or will it be used to track protesters and suppress speech? Will it transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means? Given the rise of Big Data as a socio-technical phenomenon, we argue that it is necessary to critically interrogate its assumptions and biases. In this article, we offer six provocations to spark conversations about the issues of Big Data: a cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis, and mythology that provokes extensive utopian and dystopian rhetoric.

Summary. The authors begin by acknowledging that the big data is here to stay. However, in their opinion big data is a poor choice of terminology; big data is less about sheer volume then it is about how big data can be searched, aggregated and cross referenced. The authors bring up one surprising phenomenon not usually associated with big data; they believe that big data is achieving a sort of mythological notoriety that big data can bring about higher truths than were previously possible. The authors describe six inherent weaknesses of big data as it is currently practiced:

- **The belief that big data changes the definition of knowledge**. The authors believe that big data is viewed as an inviolate source of truth and that more traditional methods for arriving at truth have lost credibility.
- **Claims to objectivity and accuracy are misleading**. The authors argue that big data can be subjective due to data interpretation and preprocessing and inaccurate due to faulty data and correlations.
- **Bigger data are not always better data**. Certain data sets, while huge, may only sample a certain population. For example, Twitter tweets only contain data for the Tweeting public. Also, combining big data sets into still larger sets may yield poorer data due to a combining of errors from the individual data sets.
- **Taken out of context, big data loses its meaning**. Historically researchers tend to focus on an individual's personal network. Today, an individual's network

includes personal contacts as well as acquaintances, co-workers, as well as strangers and also people who are brought together via communication channels, proximity to one another, and through social media interactions.

- **Just because it is accessible does not make it ethical**. Big data is very difficult to fully make anonymous. Without explicit consent from individuals in the big data set, the authors question if the use of big data is ethical in all cases.
- **Limited access to big data creates new digital divides**. Access to big data is in many cases available to those who can afford to store and use it. Smaller researchers, institutions, and businesses are at a disadvantage.

Credibility. Danah M. Boyd is a senior researcher at Microsoft Research. Boyd holds a doctorate from the School of Information (iSchool) at the University of California, Berkeley. Kate Crawford is associate professor in the Journalism and Media Research Centre at the University of New South Wales, Australia. Articles published in ICS are peer reviewed.

Jacobs, A. (2009) The pathologies of big data. *Communications of the ACM, 52*(8). Retrieved June 1, 2012 from http://queue.acm.org/detail.cfm?id=1563874.

Abstract. What is “big data” anyway? Gigabytes? Terabytes? Petabytes? A brief personal memory may provide some perspective. In the late 1980s at Columbia University I had the chance to play around with what at the time was a truly enormous “disk”: the IBM 3850 MSS (Mass Storage System). The MSS was actually a fully automatic robotic tape library and associated staging disks to make random access, if not exactly instantaneous, at least fully transparent. In Columbia’s configuration, it stored a total of around 100 GB. It was already on its way out by the time I got my hands on it, but in its heyday, the early to mid-1980s, it had been used to support access by social scientists to what was unquestionably "big data" at the time: the entire 1980 U.S. Census database.

Summary. The author first begins by stating that improvements in computer hardware such as disk storage have increased the amount of data that can be stored. The problem arises in the performance of other components of a big data application. The author creates a sample data set composed of illustrative data representing statistics about every person in the world and is able to fit the data set into a modest personal computer. Next, the author tries this with the popular database PostgreSQL and immediately runs into problems with the database. This (and most) database is optimized to handle small data queries yet the analysis of big data requires operations such as sequential processing. The author states that traditional databases are designed to load data efficiently but due to the way databases store data it is much more difficult to get data out. The author explains that for analytics the data warehouse is used to optimize data. However, the data warehouse is not designed to handle big data, either. Applications such as Excel and the statistic language R are also not up to the task of analyzing big data because they put limits on data sizes (Excel) or do not analyze efficiently in a distributed computing environment (R). The author believes that the solution to the problems of big data is distributed computing. However, drawbacks to big data in a distributed computing environment include (a) performance costs due to networking overhead and (b) the data sharding problem.

Credibility. Adam Jacobs is a senior software engineer at 1010data Inc., where, among other roles, he leads the continuing development of Tenbase, the company's ultra-high-performance analytical database engine. He holds a doctorate in neuroscience from University of California, Berkeley. The ACM publishes peer-reviewed articles. This work has been cited by 21 authors.

Manovich, L. (2011). Trending: The promises and the challenges of big social data [Online]. In M. Gold (Editor), *Debates in the Digital Humanities*. Minneapolis , MN : The University of Minnesota Press. Retrieved June 11, 2012 from http://www.manovich.net/DOCS/Manovich_trending_paper.pdf .

Abstract. The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing genetic sequences, social media interactions, health records, phone logs, government records, and other digital traces left by people. Significant questions emerge. Will large-scale search data help us create better tools, services, and public goods? Or will it usher in a new wave of privacy incursions and invasive marketing? Will data analytics help us understand online communities and political movements? Or will it be used to track protesters and suppress speech? Will it transform how we study human communication and culture, or narrow the palette of research options and alter what 'research' means? Given the rise of Big Data as a socio-technical phenomenon, we argue that it is necessary to critically interrogate its assumptions and biases. In this article, we offer six provocations to spark conversations about the issues of Big Data: a cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis, and mythology that provokes extensive utopian and dystopian rhetoric.

Summary. The author first states that the quantity of big data sets in the digital humanities is much smaller than in hard sciences, but the author believes that digital humanity big data will grow significantly due to the proliferation of online digital content. Traditionally, social scientists would sample a population and infer general trends from that population sample. This is now changed with today's interconnected networks where data streams can link individuals into interconnected networks where actual behavior is recorded. However, the author then lists four objections to the promise of big data for social understanding:

- **Only social media companies have access to really large social big data sets**. Only social scientists employed by social media companies such Google and Facebook have access to the entire big data. All other researchers rely on what big data that is exposed publically by these social media companies.
- **Many social network users build fake or misleading public personas**. Users will present a version of themselves that is carefully edited for a variety of purposes. Therefore, the user data can be misleading if used for research purposes.
- **Data depth does not equate to a deep understanding of the data.** Ethnographers gather more meaningful conclusions from their data than do data scientists. Data scientists work with great data depth but the understanding about the data is not the same. Perhaps in the future data scientists will be able to approach what ethnographers can do with machine learning.

- **The tools and expertise required to analyze big data are not available to the social and humanities researchers.** Most researchers do not have the background in computer sciences, statistics, and data mining that is required to analyze big data. Researchers who choose to use of big data directly will need to acquire skills in the use of big data for research. Also, the tools provided by social media companies are incomplete and not up to the task.

Credibility. Lev Manovich is a teacher of new media and visiting professor at the California Institute of the Arts. Manovich holds a doctorate in Visual and Cultural Studies from the University of Rochester and is the author of one book and numerous articles, presentations, and films. The ACM publishes peer-reviewed articles. This work is cited by 10 authors.

Shah, S., Horne, A., & Capella, J. (2012, April). *Good data won't guarantee good decisions*. Harvard Business Review. Retrieved June 11, 2012 from http://hbr.org/2012/04/good-data-wont-guarantee-good-decisions/ar/1?conversationId=973669.

Abstract. Global businesses have entered a new era of decision making. The ability to gather, store, access, and analyze data has grown exponentially over the past decade, and companies now spend tens of millions of dollars to manage the information streaming in from suppliers and customers. For all the breathless promises about the return on investment in Big Data, however, companies face a challenge. Investments in analytics can be useless, even harmful, unless employees can incorporate that data into complex decision-making.

Summary. The authors state that business investment in big data can be useless or even harmful if big data is not incorporated properly into complex decision-making. Through a survey of top executives, it is the opinion of the Corporate Executive Board that the best decision maker is the *informed skeptic* – this executive combines data analysis with gut feel and opinions of others to make decisions. Unfortunately the Corporate Executive Board determined that only 50% of executives are informed skeptics. The authors identify four problems that prevent the good usage of big data: (a) analytic skills are concentrated in very few employees, (b) IT spends too much time on technology and not enough time on Information, (c) reliable information is difficult to locate, and (d) business executive do not use big data because they feel they lack the expertise to do so. In order to develop more informed skeptics, business management needs to do three things: (a) increase training in data literacy, (b) incorporate data analytics into decision making, and (c) provide the right data tools.

Credibility. Shvetank Shah is executive director of The Corporate Executive Board, a management consultant company. Andrew Horne and Jaime Capella are managing directors of The Corporate Executive Board. The Harvard Business Review publishes peer-reviewed articles.

Simon, P. (2010). *Why new systems fail*. Boston, MA: Course Technology, a part of Cengage Learning.

Abstract. A Fortune 500 manufacturing company spent millions attempting to implement a new enterprise resource planning (ERP) system. Across the globe, a 150-employee marketing firm built and tried to implement a proprietary customer relationship management (CRM) system. For two very different companies doing two very different things, the outcomes were identical. In each case, the organization failed to activate and utilize its system as initially conceived by senior management. And these two organizations are hardly alone. On the contrary, research indicates that more than three in five new IT projects fail. Many miss their deadlines. Others exceed their initial budgets, often by ghastly amounts. Even systems activated on time and under budget often fail to produce their expected results and almost immediately experience major problems. Although the statistics are grim, there is at least some good news: these failures can be averted. Organizations often lack the necessary framework to minimize the chance of system failure before, during, and after beginning IT projects.

Summary. The author begins with a sobering statistic: three of five IT projects do not meet expectations of cost and performance. Also of import, the author describes four project failure modes: (a) the *unmitigated disaster*, where the project fails completely resulting in severed business relationships and law suits (uncommon); (b) the *big failure,* where the project is significantly over budget with major functionality removed; (c) the *mild failure*, where the project is slightly over budget with some reduced functionality; and (d) the *forthcoming disaster*, where some flaw in the IT project or the business organization jeopardizes the usability of the project. In order to understand the sources of system failure, the author develops themes in five parts:

Part I: Deciding to take the plunge. The author presents themes on (a) why businesses maintain legacy systems, (b) why businesses choose to implement a new system, and (c) how businesses design a replacement system.

Part II: System selection. Once a business decides to initiate a project the author presents key ideas on: (a) understanding the role of an externals sales person, (b) business process analysis and re-engineering, (c) building and maintaining support for the new system, and (d) selecting consultants.

Part III: System integration. Once the project is funded and underway the author presents key ideas on: (a) implementation strategy, (b) project communication, (c) system testing, (d) personnel roles and responsibilities, (d) reporting, and (e) documentation.

Part IV: The brave new world of post-production life. Once the project goes live the author identifies key ideas on how to achieve the projects long-term objectives: (a) ongoing system maintenance, and (b) mitigating operational changes and risks.

Part V: Maximizing the chance of success. During the duration of the project the author identifies measures that can be taken to increase the chance of success (or an orderly project termination): (a) mid-implementation corrective mechanisms, (b) audits, (c) contingency planning, (d) employee and consultant retention strategies, and (e) providing for future expansion.

Finally, the author states that in today's business many of the IT functions are changed – social media is present, business intelligence is increasingly utilized, and cloud computing is transforming the data center are among examples. However, in spite of all the changes the author believes that success of an IT project rests on people and not the technology. It is the opinion of this researcher that this work will be of value to frame the limitations of big data in the business.

Credibility. Phil Simon is an independent consultant, speaker, writer and blogger on the integration of technology in business. Simon has over 30 clients in a variety of industries including health care, public sector, and manufacturing. Course Technology publishes books on professional technical development.

Estimated Costs for the Implementation of Big Data Systems

All of the components and personnel required to implement big data need to be carefully considered by business executives and especially the CIO and CTO. Central to these executive roles in business is the concept of alignment of IT to business (Lutchen, 2004, pp. 8-10). Bantleman (2012, April 16) warns that the biggest costs in big data are in the integration of big data into the existing IT infrastructure. Trelles et al. (2011) warn that expensive data storage hardware costs are a gating cost.

Berman, K., Knight, J., & Case, J. (2008). *Financial intelligence for IT professionals: What you really need to know about the numbers* [Kindle edition]. Boston, MA: Harvard Business School Press.

Abstract. We have worked with thousands of employees, managers, and leaders in American companies, teaching them about the financial side of business. Our philosophy is that everyone in a company does better when they understand how financial success is measured and how they have an impact on the company's performance. Our term for that understanding is financial intelligence. Greater financial intelligence, we've learned, helps people feel more involved and committed. They understand better what they are a part of, what the organization is trying to achieve, and how they affect results. Trust increases, turnover decreases, and financial results improve.

Summary. The authors state begin by stating that IT leadership typically has a poor grasp of finance due the fact that IT leaders most often rise from IT departments. And yet often the CIO reports to the chief financial officer or at least must work closely with the financial organization. The authors believe that in order to be financially intelligent the CIO or CTO needs to be competent in four areas: (a) foundational (balance sheet, income statement, cash flow statement); (b) the art of making estimates, applying rules and using assumptions; (c) analyses such return on investment and financial ratios; and (d) external forces such as the economy, regulations, competition, and customer needs. In part one of the book the authors provide some guidance towards raising financial awareness: (a) that numbers can be accidentally or deliberately misleading; (b) that bias, estimates, and assumptions are encountered frequently and are presented as fact; and (c) spending the time to increase financial knowledge is key to improving a CIO's or CTO's worth to the business. Parts two through eight of the book are concerned with more detailed financial topics: (a) the income statement, (b) the balance sheet, (c) cash, (d) business ratios, (e) return on investment, (f) working capital, and (f) creating a financially savvy IT organization.

Credibility. Karen Berman is the founder and co-owner of the Business Literacy Institute. She is an expert in business and financial literacy. Berman has a doctorate in organizational psychology from the California School of Professional Psychology. Joe Knight is co-owner of the Business Literacy Institute and is also a professional speaker, trainer and author. Joe Case is a professional author. Harvard Business Publishing offers quality publications and training materials for business education. This book is cited by 3 authors.

Dumbill, E. (2012). *Planning for big data* [Kindle edition]. Santa Clara, CA : O'Reilly.

Abstract. In an age where everything is measurable, understanding big data is an essential. From creating new data-driven products through to increasing operational efficiency, big data has the potential to make your organization both more competitive and more innovative. As this emerging field transitions from the bleeding edge to enterprise infrastructure, it's vital to understand not only the technologies involved, but the organizational and cultural demands of being data-driven. Written by O'Reilly Radar's experts on big data, this anthology describes:

- The broad industry changes heralded by the big data era
- What big data is, what it means to your business, and how to start solving data problems
- The software that makes up the Hadoop big data stack, and the major enterprise vendors' Hadoop solutions
- The landscape of NoSQL databases and their relative merits
- How visualization plays an important part in data work

Summary. The author begins by stating that five years prior to the book's writing, only the largest businesses such as Walmart could afford to invest in big data. However, this has changed due to (a) open-source software such as Hadoop, (b) commodity-priced computer hardware, and (c) cloud computing. The author further develops this idea by offering book sections on (a) Hadoop, (b) a big data market survey, (c) cloud computing big data implementations, (d) commercially available big data sets, (e) NoSQL databases and (f) data analysis using visualization. Finally, the author states that big data is established as a business priority. The author states that more needs to be done to increase big data usability such as (a) better analytic tools, (b) improved data handling, and (c) the rise of big data markets.

Credibility. Edd Dumbill is program chair for O'Reilly Media's Strata and Open Source Convention conferences. Dumbill is also a technologist, a writer of four books, a blogger, and an open-source software developer. O'Reilly publishes peer-reviewed books on computer technology.

Leverich, J., & Kozyrakis, C. (2010, January). On the energy (in)efficiency of Hadoop clusters [Newsletter]. *ACM SIGOPS Operating Systems Review, 44*(1), pp. 61-65. doi: 10.1145/1740390.1740405.

Abstract. Distributed processing frameworks, such as Yahoo!'s Hadoop and Google's MapReduce, have been successful at harnessing expansive datacenter resources for large-scale data analysis. However, their effect on datacenter energy efficiency has not been scrutinized. Moreover, the file system component of these frameworks effectively precludes scale-down of clusters deploying these frameworks (i.e. operating at reduced capacity). This paper presents our early work on modifying Hadoop to allow scale-down of operational clusters. We find that running Hadoop clusters in fractional configurations can save between 9% and 50% of energy consumption, and that there is a tradeoff between performance energy consumption. We also outline further research into the energy-efficiency of these frameworks.

Summary. The authors state that energy consumption is a large component of operating a data center that hosts big data, either for private businesses or for cloud computing customers. Energy efficiency is controlled in a parallel computing environment in two ways: (a) matching the number of active nodes to the current work load, and (b) matching each nodes power consumption to its own workload. Unfortunately, MapReduce resources must remain in the active state in order to function properly, leading to great inefficiency. Also, MapReduce resources are underutilized which results in a resource that is active but mostly idle. The authors state that Hadoop has the potential to manage its resources in an energy efficient manner. A major opportunity is to turn off some nodes that are part of Hadoop's data replication. The authors discovered that if *n* nodes are involved in data replication at most *n-1* nodes can be turned off. This can result in significant savings. Finally the authors list some other factors that can be addressed to reduce costs: (a) the data layout, (b) data availability, (c) reliability and durability, (d) dynamic software job scheduling policies, (e) node architecture, and (e) workloads and applications.

Credibility. Jacob Leverich is employed by Hewlett-Packard Labs and is a doctoral candidate at Stanford University. Christos Kozyrakis is associate professor of Electrical Engineering and Computer Science at Stanford University. The ACM publishes peer-reviewed articles. This work is cited by 70 authors.

Patil, D. (2012). *Building data science teams: The skills, tools, and perspective behind great data science groups* [Kindle book]. Santa Clara, CA: O'Reilly.

Abstract. As data science evolves to become a business necessity, the importance of assembling a strong and innovative data teams grows. In this in-depth report, data scientist DJ Patil explains the skills, perspectives, tools and processes that position data science teams for success. Topics include: What it means to be "data driven", the unique roles of data scientists, the four essential qualities of data scientists, and Patil's first-hand experience building the LinkedIn data science team.

Summary. The author states that the data science profession is now emerging as critical as businesses seek to turn big data into internal products that can be used to generate revenue and growth. Data scientists are utilized in the following roles: (a) decision science and business intelligence, (b) product and marketing analysis, (c) risk management and security, (d) data services and operations, and (e) data engineering and infrastructure. The author believes that a successful data scientist is not necessarily a computer scientist or statistician but is someone who (a) possesses a technical expertise in some scientific discipline, (b) has curiosity, (c) enjoys using data to tell a story and communicates this story well, and (d) possesses cleverness to look at problems in novel ways. In order to find the best candidates, the author asks the following behavioral questions in evaluating a data scientist candidate for hire: (a) would the candidate be qualified to do a business startup, (b) can the candidate build something of immediate value to the business in 90 days, and (c) does the candidate possess the ability to do something amazing in four to six years. Finally, the author states that businesses are always trying to innovate, and the opportunities to use big data as innovation are substantial. However, the author cautions that building data science teams to tackle big data requires skill sets that are unlike teams that businesses have built before. The author states that it takes a cross-disciplinary group of very talented individuals, and finding and maintaining these teams is great challenge.

Credibility. DJ Patil is Data Scientist in Residence at Greylock Partners. Patil has held positions in academic, industry and government including the role of Chief Scientist at LinkedIn and research faculty member at the University of Maryland. Patil has a doctorate in Applied Mathematics from the University of Maryland. O'Reilly publishes peer-reviewed books on computer technology. This work is cited by 1 author.

Russom, P. (2011, 4th Quarter). Big Data Analytics. *TDWI Best Practices Report*. Retrieved June 7, 2012 from http://www.cloudtalk.it/wp-content/uploads/2012/03/1_17959_TDWIBigDataAnalytics.pdf.

Abstract. Oddly enough, big data was a serious problem just a few years ago. When data volumes started skyrocketing in the early 2000s, storage and CPU technologies were overwhelmed by the numerous terabytes of big data—to the point that IT faced a data scalability crisis. Then we were once again snatched from the jaws of defeat by Moore's law. Storage and CPUs not only developed greater capacity, speed, and intelligence; they also fell in price. Enterprises went from being unable to afford or manage big data to lavishing budgets on its collection and analysis. Today, enterprises are exploring big data to discover facts they didn't know before. This is an important task right now because the recent economic recession forced deep changes into most businesses, especially those that depend on mass consumers. Using advanced analytics, businesses can study big data to understand the current state of the business and track still-evolving aspects such as customer behavior. If you really want the lowdown on what's happening in your business, you need large volumes of highly detailed data. If you truly want to see something you've never seen before, it helps to tap into data that's never been tapped for business intelligence (BI) or analytics. Some of the untapped data will be foreign to you, coming from sensors, devices, third parties, Web applications, and social media. Some big data sources feed data unceasingly in real time. Put all that together, and you see that big data is not just about giant data volumes; it's also about an extraordinary diversity of data types, delivered at various speeds and frequencies.

Summary. The author defines the use of analytics in big data as primarily a mission to discover new facts from big data that a business can use for advantage. The time is right, the author explains, to exploit big data and analytics because: (a) big data provides huge statistical sample sizes which increase the likelihood of good results, (b) tools are now available that can handle big data, (c) the computing and networking required to process big data is affordable, (d) big data analytics can tolerate imprecise data, (e) big data analytics can allow businesses to leverage scarce resources and to fundamentally change the business. The benefits that a business derives from engaging in big data analytics include improvements in: (a) customer engagement, (b) business intelligence, and (c) specific applications like fraud detection. The barriers that a business faces in engaging in big data analytics include (a) staffing, (b) executive support, and (c) database software. The author includes in this report on big data analytics details regarding (a) organizational issues, (b) best practices, (c) tools, techniques and trends, and (d) available vendor products. Finally, the author offers recommendations on big data analytics: (a) business should embrace big data analytics for its potential; yet (b) know the barriers to big data analytics, understand the data and data-growth patterns, know hardware technology, and know the types of analytics available; and (c) continue to support the existing data warehouse and business intelligence systems.

Credibility. Philip Russom is director of TDWI research for data management. Russom is an expert in data warehousing and business intelligence. TDWI is a non-partisan organization dedicated to educating business and IT professionals about data warehousing and business intelligence, and it publishes works concerning best practices in these fields. This report is cited by 3 authors.

Conclusion

The CIO and the CTO are today being tasked with building big data into their current operations to drive significant growth in their business operations (Manyika et al., 2011). The availability of big data to business has been fueled by a significant new shift into massive data collection and analysis - what Gray (2007) calls the Fourth Paradigm. This shift is causing businesses to realign their IT operations to exploit the potential of big data while still maintaining existing operations (Schreyogg & Kliesch-Eberl, 2007). This scholarly bibliography presents and summarizes key aspects of big data that are of interest to the CIO and the CTO by examining 32 significant works in big data. IT in many businesses has been engaging in data-intensive computing for decades (Ratner, 2011), and data-intensive computing becomes big data in the presence of vast quantities of sensor network and transaction processing data (Bryant et al., 2008). In order to successfully deploy big data, the CIO and the CTO need to understand the current opportunities and limitations of big data in business. Additionally, developing costing for big data is challenging for the CIO and the CTO as the technology and staffing requirements differ significantly from traditional IT endeavors (Dumbill, 2012; Patil, 2012).

Key Factors for a CIO or CTO to Consider when Integrating Big Data into the Business

The executive role. The role of CIO with respect to the corporate suite has been a rocky one. As noted by Lutchen (2004), traditionally the CIO has been viewed as a one of a technologist – speaking in jargon that other executives could not understand; yet these other executives appreciated the importance of IT in the business and approved funding of IT at higher and higher and higher levels. The IT bubble burst of 2000 caused businesses to demand that IT become aligned with business goals and yet still provide IT services such as file sharing, email and data warehouses (Lutchen, 2004, pp. 3-5). Broadbent and Kitzis (2005) recognize these dual demands. They identify *supply side leadership* as the leadership that the CIO is charged to provide, which typically involves providing IT services. They identify *demand side leadership* as that leadership that is provided to deliver IT services concerned with creating a vision and aligning IT with the business direction.

New processes and terminology. Big data is a recent by-product of the explosion of data collection and analysis capabilities that have arisen from data collected automatically from sources such as transaction processing, mobile devices, and scientific instruments. The ensuing big data has resulted in vast data sets that are mined for information in new ways. Data users now process big data for correlations and useful patterns instead of building complex models and then examining data. Also of importance to CIOs and CTOs, the data warehousing efforts of the past do not lend themselves well to examining big data. New technologies need to be mastered in order to work with big data. Warden (2011) provides a concise yet complete glossary on big data.

Implementing big data. Lutchen (2004) provides insight into the role of CIO and CTO when implementing IT within the business and writes that this role six parts. The three parts that are related to the introduction of big data are (a) big data alignment to business goals, (b) integration of big data as an emerging technology, and (c) leveraging big data in the business. The following sections focus specifically on aspects of each.

In *big data alignment* to business goals, Lutchen (2004) writes that the CIO and CTO need to directly link the efforts of IT to the business environment. Buchan and Win (as cited In Hey et al., 2009) write about big data's potential in unifying health records and biomedical research to provide new types of value to health care providers and patients. Walmart is a business that relies on big data from its transaction processing and warehouse inventory system to drive value from its vast operation (Cukier, 2010).

In *integrating big data* as an emerging technology the CIO and CTO are responsible for the deployment of technology to support big data. Agrawal, Das, and Abbadi (2011) write that big data will require (a) cloud computing resources, (b) Hadoop for data storage and organization, and (c) NoSQL data storage. Borkar et al. (2012) state that Hadoop is a good choice for big data, and the Hadoop open-source software system will continue to evolve to handle big data.

In *leveraging big data* in the business the CIO ensures that big data is useful to the business and is fully integrated into decision making. Friendly writes about the importance of visualization in big data analysis. Linoff and Berry (2011) provide data mining techniques that can be applied to big data to improve business engagement with their customers. Bollier (2010) describes how alerting can be used to great advantage in improving health care.

The Nature of Data-intensive Computing and How it is Related to Big data

As noted by Cohen et al. (2009), the CIO and the CTO have managed data-intensive computing for decades. Traditionally, IT departments have leveraged existing databases to create data warehouses from which analytical reports could be derived for business intelligence. The data warehouse involved very lengthy development projects, and the data warehouse was very difficult to change once it was deployed into the business (Borkar et al., 2012).

Gray and Szalay (2007) saw that data-intensive computing had entered a new phase due to proliferation of big data caused by the influx of sensor networks; they called this the *fourth paradigm*. In the fourth paradigm data exploration is the primary method of inquiry, as compared to the third paradigm where computation and modeling was the norm. Bell (2009) defines data-intensive computing as composed of three activities: (a) data capture, (b) data curation (organizing data in some logical manner), and (c) analysis. Each one needs to be carefully addressed when dealing with big data.

Several examples of-data intensive computing provide illustration of what can be done with big data. Ali, Chen and Lee (2008) describe how manufacturing can benefit from big data in data-intensive computing by *enabling expert systems to mine sensor networks* in an agile and flexible manufacturing environment. Ali et al. (2008) believe the net result is optimized manufacturing and reduced defects. Cohen et al. (2009) write that *new techniques of data analysis need to be used* with big data in data-intensive computing because the data warehouse model no longer applies to big data. Kouzes et al. (2009) write that a broader definition of data-intensive is required with big data. Kouzes et al. (2009) believe that data-intensive computing is *a revolutionary shift away from data warehousing to applications* that are concerned with the rapid processing of data streams to achieve timely and useful business analytics.

In the conclusion to *The fourth paradigm: Data-intensive scientific discovery* Hey et al. (2009) write that data-intensive computing is now requiring multi-disciplinary efforts in order to solve the problems of big data. The CIO and CTO need to be aware of this. Also the fact that the growth of funding for research has not kept pace with the introduction of big data means that the CIO and CTO need to be aware of new developments, so that data-intensive computing in their business adapts to improvements in big data technology.

The Current Opportunities of Big Data Applications

Timeliness to engage in big data. Bryant et al. (2008) believe that now is the ideal time for the CIO and CTO to engage IT in big data. According to Manyika et al. (2011), big data is currently being integrated into business due to the emergence of technology like Hadoop and cloud computing. While some business sectors such as technology are ahead of the curve in their usage of big data, other sectors such as health care, manufacturing and the public sector are not yet on board. This is problematic, since these types of organizations are facing cost pressures, and big data can be incorporated in business operations to reduce costs and streamline operations (Manyika et al., 2011). Ratner (2011) states that customer-relationship management is a key cross-business-sector goal, and he describes how the CIO and CTO can use big data analytics to improve customer retention.

Availability of suitable technology. Bryant et al. (2008) write the promise of big data is made possible by the integration of key technology such as sensors, data storage, distributed computing, and analytics. The CIO and CTO need to spend significantly on these capabilities. However, Bryant et al. (2008) caution that big data technology is still emerging, and the CIO and CTO need be careful not to overinvest in technology that may quickly become obsolete.

Analytics and machine learning maturation. Analytics refers to the output product of big data (Bell, 2009), and the CIO and CTO need to ensure that big data analytics are available and tightly integrated into the business (Franks, 2012). LaValle, Lesser, Shockley, Hopkins and Kruschwitz (2010) report that top business performers are five times more likely to use analytics then the lowest business performers. Eagle (2010) reports that analytics can be used derive inferences about large population segments by sampling very small population sizes and then correlating the samples to the large population. Witten, Elbe and Hall (2011) write that analytics can be combined with machine learning to produce rich data mining opportunities for the business (Langley (1996) describes *machine learning* as providing machines with the ability to acquire intelligence through automatic means).

Integration of big data with existing IT infrastructure. Schreyogg and Kliesch-Eberl (2007) believe that big data is not an all-or-nothing proposition. The CIO and the CTO need to ensure that big data complements and enhances the existing IT infrastructure. As an example, Schreyogg and Kliesch-Eberl (2007) propose that big data can be used as an early-warning alert system to monitor existing business processes.

The Current Limitations of Big Data Systems

High percentage of project failure. Simon (2010) provides a sobering statistic: three of five projects do not meet expectations for cost and performance. While some projects are complete failures at the outset, other projects fail for unexpected reasons long after the project has been deployed. The CIO and the CTO need to guard against expensive IT failures in the business by recognizing and guarding against encountering the limitations of big data.

Integrate big data and business functions. The CIO and the CTO should engage in big data projects for well-documented reasons (Manyika et al., 2011). Once the big data project is underway, limitations are encountered such as how tightly to integrate big data into the existing business infrastructure (Schreyogg & Kliesch-Eberl, 2007). Shah, Horne, and Capella (2012, April) write that big data integration can be hampered from being fully integrated in the business if too few decision makers within the business are trained in the use of big data tools such as analytics. Also, Shah et al. (2012, April) write that the CIO and the CTO need to guard against too much emphasis being placed on technology and not enough emphasis on business integration.

Proper hardware selection. Simon (2010) writes that system selection is another area that the CIO and the CTO will find to be a limitation. Jacobs (2009) writes that hardware selection is crucial to the success of big data; he cites that distributed computing is a requirement to analyzing big data, yet limitations are imposed by networking. Another major limitation that can occur is when an aggregated data subset in a distributed system needs to be accessed frequently by other data subsets. The resultant throughput delays can cause inefficiencies in distributed computing (Jacobs, 2009).

Staff selection and retention. Staff selection and retention is another area that Simon (2010) says will pose a challenge. In particular, Manyika et al. (2011) write about the critical shortage in data scientists. The CIO and the CTO must ensure that staffing is adequate and that personnel retention efforts are in place.

Cautious data interpretation. Finally, Simon (2010) writes that there are post-production limitations to what can be done with big data. Boyd and Crawford (2012) concur that big data has inherent weaknesses in its worth such as misleading data and data weakness when taken out of a specific context; they also state that limitations imposed by weak data correlation are frequently present. As an example, Manovich (2011) reports that big data provided by social media companies can be misleading because they are only providing a subset of their entire data set.

The Cost Estimation of Big Data Systems

The CIO and the CTO need to be competent in four areas of finance, according to Berman, Knight and Case (2008): (a) the foundational area (balance sheet, income statement, cash flow statement); (b) the art of making estimates, applying rules and using assumptions; (c) analyses, such as return on investment and financial ratios; and (d) external forces, such as the economy, regulations, competition, and customer needs (Preface, para. 9).

In the foundational area, Berman et al. (2008) note that the CIO and the CTO have direct input via operating costs. For example, Leverich and Kozyrakis (2010, January) state that data centers running Hadoop can be very costly to operate, and they identify ways in which the costs of running Hadoop can be reduced – some ways are achieved by properly configuring Hadoop, while others require data center hardware improvements.

Concerning the art of making estimates, Berman et al. (2008) note that the CIO and the CTO have to rely on rule of thumb, assumptions, and experience in order to provide justification for expenditures in big data. Unfortunately, big data is too new and therefore accurate business financial data is not available. Dumbill (2012) writes that many commercial big data components such as analytics and big data sets are available. The CIO and the CTO can prepare estimates for big data implementation that can be presented for executive approval.

In order to understand external forces, Berman et al. (2008) note that the CIO and CTO must rely on a skilled and well-trained IT staff to be able to assess how external forces – regulation, security, and user behavior – affect how successful big data becomes as a transformative force in the business. Patil (2012) writes about the need to hire, develop, and retain IT professionals including data scientists; he states strongly that the best data scientists are extremely hard to find and retaining a business's data scientists is crucial.

Through carefully planned financial analysis, Berman et al. (2008) note that the CIO and the CTO can prepare formal analysis that can be used to justify big data expenditures. Analytics is almost always the main focus of big data, and Russom (2011, 4th Quarter) describes the components that make up the big data system. If the CTO and CTO cost the components identified by Russom (2011, 4th Quarter) and then apply estimates as to how much big data could save or make for a business, then project justification metrics such as return on investment can be calculated.

References

A guide to evaluating resources. (n.d.). Retrieved June 11, 2012 from http://www.cornellcollege.edu/politics/courses/allin/Misc/guide%20to%20evaluating%20resources.pdf.

Agrawal, D., Das S., & Abbadi, A. (2011). Big data and cloud computing: Current state and future opportunities. *Proceedings of EDBT*, 530-533. Retrieved June 2, 2012 from http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf.

Ali, A., Chen, Z., & Lee, J. (2008). Web-enabled platform for distributed and dynamic decision-making systems. *International Journal of Advanced Manufacturing Technology, 38*, 11-12. Retrieved June 1, 2012 from http://www.springerlink.com/content/n2333763960716q6/.

Bantleman, J. (2012, April 16). The big cost of big data. In E. Savitz, *CIO network: Insights and ideas for technology leaders* [Web log post]. Forbes Magazine. Retrieved June 11, 2012 from http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/.

Bargmeyer, B., & Gillman, D. (n.d.). Metadata standards and metadata registries: An overview. Bureau of Labor Statistics. Retrieved June 5, 2012 from http://www.bls.gov/ore/pdf/st000010.pdf.

Barney, B. (n.d.). What is parallel computing? *Introduction to Parallel Computing*. Retrieved June 11, 2012 from https://computing.llnl.gov/tutorials/parallel_comp/#Whatis.

Bastek, N., Robinson, J., & Barnes, L. (2012). *What is a review paper?* Retrieved on May 1, 2012 from http://writing.colostate.edu/guides/documents/review_essay/pop2a.cfm.

Beel, J., Gipp, B., & Stiller, J. (2009). Information Retrieval on Mind Maps – What could it be good for? *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09)*, 1–4. Retrieved June 11, 2012 from http://sciplore.org/wp-content/papercite-data/pdf/beel09f.pdf.

Bell, C., & Smith, T. (2009). *Critical evaluation of information sources.* Retrieved June 11, 2012 from http://libweb.uoregon.edu/guides/findarticles/credibility.html.

Bell, G. (2009). Foreword. In T. Hey, S. Tansley & K. Tolle. *The fourth paradigm: Data intensive scientific discovery*. Redmond, WA: Microsoft Research. Retrieved June 11, 2012 from http://research.microsoft.com/en-us/collaboration/fourthparadigm/contents.aspx.

Bell, G., Hey, T., & Szalay, A. (2009, March). Beyond the data deluge. *Science,* 1297-1298. Retrieved June 11, 2012 from http://www.sciencemag.org/content/323/5919/1297.summary.

Berman, K., Knight, J., & Case, J. (2008). *Financial intelligence for IT professionals: What you really need to know about the numbers* [Kindle edition]. Boston, MA: Harvard Business School Press.

Big data. (n.d.). *Wikipedia*. Retrieved May 1, 2012 from http://en.wikipedia.org/wiki/Big_data.

Big data bibliography – compiled by the Safari Books Online Content Team [Kindle edition]. (2011). O'Reilly.

Big data conference [Web site]. (2012). Retrieved from http://www.bigdataconference.net.

Bollier, D. (2010). *The promise and peril of big data*. Washington, DC: The Aspen Institute. Retrieved June 11, 2012 from https://www.c3e.info/uploaded_docs/aspenbig_data.pdf.

Borkar, V., Carey, M., & Li, C. (2012). Inside big data management: Ogres, onions, or parfaits? *EDBT*. Retrieved April 29, 2012 from http://www.edbt.org/Proceedings/2012-Berlin/papers/keynotes/a3-carey.pdf.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication and Society,* 662-679. doi: 10.1080/1369118X.2012.678878.

Broadbent, M., & Kitzis, E. (2005). *The new CIO leader: Setting the agenda and delivering results*. Boston, MA: Havard Business School Press.

Brownstein, J., Freifeld, C., & Madoff, L. (2009). Influenza A (H1N1) virus, 2009 — online monitoring. *New England Journal of Medicine*, *360*(21), 2156. doi: 10.1056/NEJMp0904012.

Bryant, R. E., Katz, R. H., & Lazowska, E. D. (2008). *Big-data computing: Creating revolutionary breakthroughs in commerce, science and society*. Computing Research Association. Retrieved June 11, 2012 from http://www.cra.org/ccc/docs/init/Big_Data.pdf.

Budd, J. (2004). Mind maps as classroom exercises. *The Journal of Economic Education*, *35*(1).

Busch, C., De Maret, P., Flynn, T., Kellum, R., Le, S., Meyers, B., Saunders, M., White, R., & Palmquist, M. (2005). Content analysis. Writing@CSU. *Colorado State University Department of English*. Retrieved May 1, 2012 from http://writing.colostate.edu/guides/research/content/.

Cannataro, M., Talia, D., & Srimani, P. (2002). Parallel data-intensive vomputing in scientific and commercial applications. *Parallel Computing,* 673-704.

Carrie, C. (2006). The new electronic trading regime of dark books: Mashups and algorithmic trading. *Institutional Investor Journals 2006*, 1, 14-20.

Cohen J., Dolan, B., Dunlap, M., Hellerstein, J. M., & Welton, C. (2009). MAD skills: New analysis practices for big data. *Proceedings of the VLDB Endowment, 2*(2), 1481-1492. Retrieved June 1, 2012 from http://dl.acm.org/citation.cfm?id=1687553.1687576.

Copoulos, M. (2012, April). 10 key takeaways on the stage 2 meaningful-use proposal. *Health Management Technology*, 4-5.

Correlation. (2012). In *Encyclopædia Britannica*. Retrieved June 25, 2012 from http://www.britannica.com/EBchecked/topic/690049/correlation .

Creswell, J. (2009). *Research design: qualitative, quantitative, and mixed method approaches (3rd ed.)*. Thousand Oaks, CA: Sage Publications, Inc.

Cukier, K. (2010). Data, data everywhere. *The Economist*. Retrieved June 11, 2012 from http://www.economist.com/node/15557443?story_id=15557443.

Data mining**.** (2012). In *Encyclopædia Britannica*. Retrieved June 11, 2012 from http://www.britannica.com/EBchecked/topic/1056150/data-mining.

Distributed computing. (2012). In *Encyclopædia Britannica*. Retrieved from http://www.britannica.com/EBchecked/topic/1494997/distributed-computing.

Druce, L. (2009). *Q & A with mind mapping guru Tony Buzan*. KnowledgeBoard. Retrieved from http://www.knowledgeboard.com/item/2980.

Dumbill, E. (2012). *Planning for big data* [Kindle edition]. Santa Clara, CA: O'Reilly.

Eagle, N. (2010). Big data, global development, and complex social systems. *Proceedings of the eighteenth ACM SIGSOFT International Symposium on Foundations of Software*

Engineering (FSE '10), 3-4. Retrieved June 2, 2012 from http://www.legacyforhealth.org/PDF/Eagle.pdf.

Earl, M., & Scott, I. (1999). Opinion: What is a chief knowledge officer? *Sloan Management Review, 40*(2), 29. Retrieved June 11, 2012 from http://itu.dk/people/petermeldgaard/km/lektion%2012/1999-EarlScott.pdf.

Expert system. (2012). In *Encyclopædia Britannica*. Retrieved June 11, 2012 from http://www.britannica.com/EBchecked/topic/198506/expert-system.

Finnerty, J., & Park, H. (1988, Winter). How to profit from program trading. *The Journal of Portfolio Management, 14*(2), 40-46. doi: 10.3905/jpm.1988.409134.

Firestone, C. (2010). Foreword. In D. Bollier, *The promise and peril of big data* (pp. vii - ix). Washington, DC: The Aspen Institute. Retrieved June 11, 2012 from https://www.c3e.info/uploaded_docs/aspenbig_data.pdf.

Foster, I. (2002). *What is the grid? A Three Point Checklist.* Retrieved June 11, 2012 from http://dlib.cs.odu.edu/WhatIsTheGrid.pdf.

Franks, B. (2012). *Taming the big data tidal wave: Finding opportunities in huge data streams with advanced analytics* [Kindle edition]. Hoboken, NJ: John Wiley & Sons, Inc.

Freeplane. (n.d.). Retrieved from http://freeplane.sourceforge.net/wiki/index.php/Main_Page.

Freifeld, C., Mandl, K., Reis, B., & Brownstein. (2008). HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Information Association, 15*(2), 150-157.

Friendly, M. (2009). *Milestones in the history of thematic cartography, statistical graphics,* ***and*** *data visualization*. Retrieved June 11, 2012 from http://euclid.psych.yorku.ca/SCS/Gallery/milestone/milestone.pdf.

Gorton, I., Greenfield, P., Szalay, A., & Williams, R. (2008, April). Data-intensive computing in the 21st Century, *Computer*, *41*(4), 30-32. doi: 10.1109/MC.2008.122.

Gray, J., & Szalay,A. (2007). *eScience—A transformed scientific method.* Presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA. Retrieved from http://www.slideshare.net/dullhunk/escience-a-transformed-scientific-method.

Henshen, D. (2011, November). Why all the hadoopla? *Information Week,* 19-26.

Hey, T. (n.d.). *eScience, Semantic computing and the cloud: Towards a smart cyberinfrastructure and the cloud*. Retrieved June 11, 2012 from computerlectures.pnnl.gov/pdf/hey_presentation.pdf.

Hey, T., Tansley, S., & Tolle, K. (2009). *The fourth paradigm: Data-intensive scientific discovery* [Kindle version]. Redmond, WA: Microsoft Research.

Hey, T., & Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G. Fox & T. Hey (Eds.), *Grid Computing: Making the global infrastructure a reality*. Wiley Online Library. Retrieved June 11, 2012 from http://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf.

Hopkins, M. (2010). The 4 ways IT is revolutionizing innovation. *MIT Sloan Management Review*, *51*(3), 51-56.

Information system. (2012). In *Encyclopædia Britannica*. Retrieved June 11, 2012 from http://www.britannica.com/EBchecked/topic/287895/information-system/302406/Databases-and-data-warehouses#ref1134725.

Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM, 52*(8). Retrieved June 1, 2012 from http://queue.acm.org/detail.cfm?id=1563874.

Jha, A. (2010). Meaningful use of electronic health records: The road ahead. *The Journal of the American Medical Association*, *304*(15), 1709-1710. doi: 10.1001/jama.2010.1497.

Johnston, W.E. (1998, July). High-speed, wide-area, data intensive computing: A ten year retrospective. *7th IEEE Symposium on High Performance Distributed Computing.* Symposium conducted in Chicago, IL. Retrieved July 1, 2012 from http://acs.lbl.gov/~johnston/papers/TenYearRetrospective_IEEE2col.pdf.

Joshi, M. (2005). *Presentation for graduate course in advanced computer architecture*. Retrieved June 11, 2012 from http://www.d.umn.edu/~joshi031/files/grid-computing.pdf.

Kenwright, D. (1999). Automation or interaction: What's best for big data? *Visualization '99 Proceedings,* 491 -495. doi: 10.1109/VISUAL.1999.809940.

Koc, M., Ni, J., Lee, J., & Bandyopadhyay, P. (2004). *Introduction to e-manufacturing*. University of Michigan. Retrieved from http://wumrc.engin.umich.edu/junni/publications/Introductionofe-Manufacturing.pdf.

Kohavi, R., Rothleder, N., & Simoudis, E. (2002). Emerging trends in business analytics. *Communications of the ACM, 45*(8), 45-48.

Kouzes, R., Anderson, G., Elbert, S., Gorton, I., & Gracio, D. (2009). The changing paradigm of data-intensive computing. *Computer*, *42*(1), 26-34. doi: 10.1109/MC.2009.26.

KU Writing Center, The University of Kansas. (2011, July). *Writing guide: References and bibliographies*. Retrieved June 11, 2012 from http://www.writing.ku.edu/~writing/guides/documents/bibs.pdf.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. In *Application delivery strategies, File: 949*. Meta Group. Retrieved June 11, 2012 from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Langley, P. (1996). *Elements of Machine Learning*. San Francisco, CA: Morgan Kaufmann.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M., & Kruschwitz, N. (2010). Big data, analytics, and the path from insights to value. *MIT Sloan Management Review*, *62*(2). Retrieved June 2, 2012 from http://sloanreview.mit.edu/the-magazine/2011-winter/.

Leverich, J., & Kozyrakis, C. (2010, January). On the energy (in)efficiency of Hadoop clusters [Newsletter]. *ACM SIGOPS Operating Systems Review*, *44*(1), pp. 61-65. doi: 10.1145/1740390.1740405.

Linoff, G., & Berry, M. (2011). *Data mining techniques: For marketing, sales, and customer relationship management*. Indianapolis, IN: Wiley Publishing.

Lutchen, M. (2004). *Managing IT as a business: A survival guide for CEOs*. Hoboken, NJ: John Wiley and Sons.

Lynch, C. (2008, September). Big data: How do your data grow? *Nature*, *455*, 28-29. doi:10.1038/455028a.

Machi, L., & McEvoy, B. (2009). *The literature review: Six steps to success* [Kindle version]. Thousand Oaks, CA: Corwin Press.

Manovich, L. (2011). Trending: The promises and the challenges of big social data [Online]. In M. Gold (Editor), *Debates in the Digital Humanities* , In M. Gold. Minneapolis , MN : The University of Minnesota Press. Retrieved June 11, 2012 from http://www.manovich.net/DOCS/Manovich_trending_paper.pdf.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). *Big data: The next frontier for innovation, competition, and productivity* [Kindle edition] . McKinsey Global Institute. Retrieved June 11, 2012 from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

Mell, P., & Grance, T. (2011). *The NIST definition of cloud computing: Recommendations of the National Institute of Standards and Technology.* Retrieved June 11, 2012 from http://rszt.pmmk.pte.hu/uploads/8f/23/8f23a309550830fa62395163ecec6fd3/nist_SP800-145.pdf.

Mind map. (n.d.) *Wikipedia*. Retrieved May 10, 2012 from http://en.wikipedia.org/wiki/Mind_map.

Moore, R., Baru, C., Marciano, R., Rajasekar, A., & Wan, M. (1999). Chapter 5: Data-intensive computing. In I. Foster, C. Kesselman, *The grid: Blueprint for a new computing infrastructure.* San Francisco, CA: Elsevier (Morgan Kaufman).

Olhoff, J. (2011). *How to write a literature review* [Kindle version]. Farmington, MN: Sparrow Media Group.

Olson, M. (2011, January 12). Hadoop: What it is, how it works, and what it can do [Web log post]. In J. Turner, *Radar: News and Commentary (Data).* Retrieved June 11, 2012 from http://radar.oreilly.com/2011/01/what-is-hadoop.html.

Open source. (2012). In *Encyclopædia Britannica*. Retrieved June 5, 2012 from http://www.britannica.com/EBchecked/topic/1017825/open-source.

Pallay, J. (2005). Milliseconds matter. *Wall Street & Technology*, 28-30.

Papazoglou, M., & Georgakopoulos, D. (2006). Service-oriented computing. *Communications of the ACM, 46*(10), 25-28.

Patil, D. (2012). *Building data science teams: The skills, tools, and perspective behind great data science groups* [Kindle book]. Santa Clara, CA : O'Reilly.

Rappa, M. (2012, May 3). What is a data scientist? Michael Rappa, Institute for Advanced Analytics. In D. Woods, *Data Driven*. Forbes. Retrieved June 11, 2012 from http://www.forbes.com/sites/danwoods/2012/03/05/what-is-a-data-scientist-michael-rappa-north-carolina-state-university/2/.

Raschke, R., Gollihare, B., Wunderlich, T., Guidry, J., Leibowitz, A., Peirce, J., Lemelson, L., Heisler, M., & Susong, C. (1998, October 21). A computer alert system to prevent injury from adverse drug events: Development and evaluation in a community teaching hospital. *JAMA, 280*(15), 1317-1320. doi:10.1001/jama.280.15.1317.

Ratner, B. (2011). *Statistical and machine-learning data mining: Techniques for better predictive modeling and analysis of big data, second edition.* CRC Press: New York.

Ruiz, E., Castillo, C., Hristides, V., Gionis, A., & Jaimes, A. (2012). *Correlating financial time series with micro-blogging activity.* University of California, Riverside. Retrieved from http://www.cs.ucr.edu/~vagelis/publications/wsdm2012-microblog-financial.pdf.

Russom, P. (2011, 4th Quarter). Big Data Analytics. *TDWI Best Practices Report*. Retrieved June 7, 2012 from http://www.cloudtalk.it/wp-content/uploads/2012/03/1_17959_TDWIBigDataAnalytics.pdf.

Schreyogg, G., & Kliesch-Eberl, M. (2007). How dynamic can organizational capabilities be? Towards a dual-process model of capability dynamization. *Strategic Management Journal, 28*(9)*,* 913-933. doi: 10.1002/smj.613

Shah, S., Horne, A., & Capella, J. (2012, April). *Good data won't guarantee good decisions*. Harvard Business Review. Retrieved June 11, 2012 from http://hbr.org/2012/04/good-data-wont-guarantee-good-decisions/ar/1?conversationId=973669.

Shapira, B., Elovici, Y., Last, M., & Kandel, A. (2008). Enhancement to the Advanced Terrorist Detection System (ATD). In C. Gal, P. Kantor, P & B. Shapira (Eds.), *Security informatics and terrorism: Patrolling the web: Social and technical problems of detecting and controlling terrorist's use of the world wide web,* (pp. 71-81). Beer-Sheva: Israel: IOS Press.

Silberschatz, A., Peterson, J., & Galvin, P. (1991). Operating System Concepts (3rd ed.). Reading, MA: Addison-Wesley.

Simon, P. (2010). *Why new systems fail*. Boston, MA: Course Technology, a part of Cengage Learning.

Sloan Digital Sky Survey. (2012). Retrieved April 30, 2012 from http://www.sdss.org.

Smith, R. (2003, July-August). The chief technology officer: Strategic responsibilities and relationships. *Research Technology Management*. Retrieved June 11, 2012 from http://www.modelbenders.com/papers/SmithR_CTOStrategy.pdf.

Sokolinsky, L. (2004). Survey of architectures of parallel database systems. *Programming and Computer Software, 30*(6), 337-346. Retrieved June 11, 2012 from http://www.springerlink.com/content/r7445g1185183325/fulltext.pdf.

Szalay, A. (2010). *Data driven discovery in science: the Fourth Paradigm*. Institute for Data Intensive Engineering and Science, Johns Hopkins University. Retrieved June 11, 2012 from http://idies.jhu.edu/seminars.aspx.

Trelles, O., Prins, P., Snir, M., & Jansen, R. (2011). Big data, but are we ready? *Nature Reviews Genetics*, *12*(224). doi:10.1038/nrg2857-c1.

Warden, P. (2011). *Big data glossary* [Kindle version]. Sebastopol, CA : O,Reilly Media.

Witten, I., Elbe, F., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann.

Zotero. [Web site] (n.d.). Retrieved June 11, 2012 from http://www.zotero.org/.