Chapter 3
Big Data Generation and Acquisition

Abstract We have introduced several key technologies related to big data, i.e., cloud computing, IoT, data center, and Hadoop. Next, we will focus on the value chain of big data, which can be generally divided into four phases: data generation, data acquisition, data storage, and data analysis. If we take data as a raw material, data generation and data acquisition are exploitation process, data storage is a storage process, and data analysis is a production process that utilizes the raw material to create new value.

3.1 Big Data Generation

Data generation is the first step of big data. Specifically, it is large-scale, highly diverse, and complex datasets generated through longitudinal and distributed data sources. Such data sources include sensors, videos, click streams, and/or all other available data sources. At present, main sources of big data are the operation and trading information in enterprises, logistic and sensing information in the IoT, human interaction information and position information in the Internet world, and data generated in scientific research, etc. The information far surpasses the capacities of IT architectures and infrastructures of existing enterprises, while its real-time requirement also greatly stresses the existing computing capacity.

3.1.1 Enterprise Data

In 2013, IBM issued a reported titled "Analytics: The Real-world Use of Big Data," which indicates that the internal data of enterprises are the main sources of big data. The internal data of enterprises mainly consists of online trading data and online analysis data, most of which are historically static data and are managed by RDBMSs in a structured manner. In addition, production data, inventory data, sales

M. Chen et al., *Big Data: Related Technologies, Challenges and Future Prospects*, SpringerBriefs in Computer Science, DOI 10.1007/978-3-319-06245-7_3,

data, and financial data, etc., also constitute enterprise internal data, which aims to capture informationized and data-driven activities in enterprises, so as to record all activities of enterprises in the form of internal data.

Over the past decades, IT and digital data have contributed a lot to improve the profitability of business departments. It is estimated that the business data volume of all companies in the world may double every 1.2 years [1], in which, the business turnover through the Internet, enterprises to enterprises, and enterprises to consumers per day will reach USD 450 billion [2]. The continuously increasing business data volume requires more effective real-time analysis so as to fully harvest its potential. For example, Amazon processes millions of terminal operations and more than 500,000 queries from third-party sellers per day [3]. Walmart processes one million customer trades per hour and such trading data are imported into a database with a capacity of over 2.5PB [4]. Akamai analyzes 75 million events per day for its target advertisements [5].

3.1.2 IoT Data

As discussed, IoT is an important source of big data. Among smart cities constructed based on IoT, big data may come from industry, agriculture, traffic and transportation, medical care, public departments, and households, etc., as shown in Fig. 3.1.

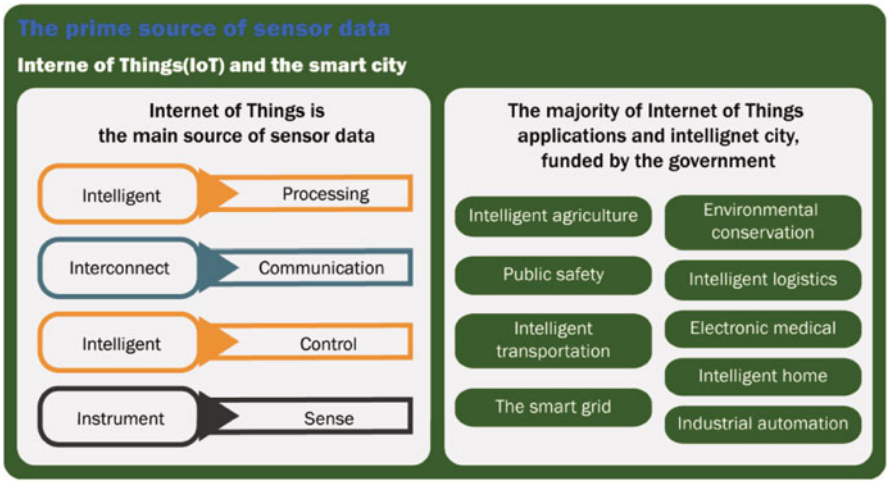

Fig. 3.1 Illustration of the prime source of sensory data

According to the processes of data acquisition and transmission in IoT, its network architecture may be divided into three layers: the sensing layer, the network layer, and the application layer. The sensing layer is responsible for data acquisition

and mainly consists of sensor networks. The network layer is responsible for information transmission and processing, where close transmission may rely on sensor networks, and remote transmission shall depend on the Internet. Finally, the application layer support specific applications of IoT.

According to the characteristics of IoT, the data generated from IoT has the following features:

- *Large-Scale Data*: in IoT, masses of data acquisition equipments are distributedly deployed, which may acquire simple numeric data (e.g., location) or complex multimedia data (e.g., surveillance video). In order to meet the demands of analysis and processing, not only the currently acquired data, but also the historical data within a certain time frame should be stored. Therefore, data generated by IoT are characterized by large scales.
- *Heterogeneity*: because of the variety data acquisition devices, the acquired data is also different and such data features heterogeneity.
- *Strong Time and Space Correlation*: in IoT, every data acquisition device are placed at a specific geographic location and every piece of data has a time stamp. The time and space correlations are important properties of data from IoT. During data analysis and processing, time and space are also important dimensions for statistical analysis.
- *Effective Data Accounts for Only a Small Portion of the Big Data*: a great quantity of noises may occur during the acquisition and transmission of data in IoT. Among datasets acquired by acquisition devices, only a small amount of abnormal data is valuable. For example, during the acquisition of traffic video, the few video frames that capture the violation of traffic regulations and traffic accidents are more valuable than those only capturing the normal flow of traffic.

3.1.3 Internet Data

Internet data consists of searching entries, Internet forum posts, chatting records, and microblog messages, among others, which have similar features, such as high value and low density. Such Internet data may be valueless individually, but through exploitation of accumulated big data, useful information such as habits and hobbies of users can be identified, and it is even possible to forecast users' behavior and emotional moods.

3.1.4 Bio-medical Data

As a series of high-throughput bio-measurement technologies are innovatively developed in the beginning of the twenty-first century, the frontier research in the bio-medicine field also enters the era of big data. By constructing smart,

efficient, and accurate analytical models and theoretical systems for bio-medicine applications, the essential governing mechanism behind complex biological phenomena may be revealed. Not only the future development of bio-medicine can be determined, but also the leading roles can be assumed in the development of a series of important strategic industries related to the national economy, people's livelihood, and national security, with important applications such as medical care, new drug R&D, and grain production (e.g., transgenic crops).

The completion of HGP (Human Genome Project) and the continued development of sequencing technology also lead to widespread applications of big data in the field. The masses of data generated by gene sequencing go through specialized analysis according to different application demands, to combine it with the clinical gene diagnosis and provide valuable information for early diagnosis and personalized treatment of disease. One sequencing of human gene may generate 100–600 GB raw data. In the China National Genebank in Shenzhen, there are 1.3 million samples including 1.15 million human samples and 150,000 animal, plant, and microorganism samples. By the end of 2013, 10 million traceable biological samples will be stored, and by the end of 2015, this figure will reach 30 million. It is predictable that, with the development of bio-medicine technologies, gene sequencing will become faster and more convenient, and thus making big data of bio-medicine continuously grow beyond all doubt.

In addition, data generated from clinical medical care and medical R&D also rise quickly. For example, the University of Pittsburgh Medical Center (UPMC) has stored 2TB such data. Explorys, an American company, provides platforms to collocate clinical data, operation and maintenance data, and financial data. At present, about 13 million people's information have been collocated, with 44 articles of data at the scale of about 60TB, which will reach 70TB in 2013. Practice Fusion, another American company, manages electronic medical records of about 200,000 patients.

Apart from such small and medium-sized enterprises, other well-known IT companies, such as Google, Microsoft, and IBM have invested extensively in the research and computational analysis of methods related to high-throughput biological big data, for shares in the huge market as known as the "Next Internet." IBM forecasts, in the 2013 Strategy Conference, that with the sharp increase of medical images and electronic medical records, medical professionals may utilize big data to extract useful clinical information from masses of data to obtain a medical history and forecast treatment effects, thus improving patient care and reduce cost. It is anticipated that, by 2015, the average data volume of every hospital will increase from 167TB to 665TB.

3.1.5 Data Generation from Other Fields

As scientific applications are increasing, the scale of datasets is gradually expanding, and the development of some disciplines greatly relies on the analysis of masses

of data. Here, we examine several such applications. Although being in different scientific fields, the applications have similar and increasing demand on data analysis. The first example is related to computational biology. GenBank is a nucleotide sequence database maintained by the U.S. National Bio-Technology Innovation Center. Data in this database may double every 10 months. By August 2009, Genbank has more than 250 billion bases from 150,000 different organisms [6]. The second example is related to astronomy. Sloan Digital Sky Survey (SDSS), the biggest sky survey project in astronomy, has recorded 25TB data from 1998 to 2008. As the resolution of the telescope is improved, by 2004, the data volume generated per night will surpass 20TB. The last application is related to high-energy physics. In the beginning of 2008, the Atlas experiment of Large Hadron Collider (LHC) of European Organization for Nuclear Research generates raw data at 2PB/s and stores about 10TB processed data per year.

In addition, pervasive sensing and computing among nature, commercial, Internet, government, and social environments are generating heterogeneous data with unprecedented complexity. These datasets have their unique data characteristics in scale, time dimension, and data category. For example, mobile data were recorded with respect to positions, movement, approximation degrees, communications, multimedia, use of applications, and audio environment. According to the application environment and requirements, such datasets can be classified into different categories, so as to select the proper and feasible solutions for big data.

3.2 Big Data Acquisition

As the second phase of the big data system, big data acquisition includes data collection, data transmission, and data pre-processing. During big data acquisition, once the raw data is collected, an efficient transmission mechanism should be used to send it to a proper storage management system to support different analytical applications. The collected datasets may sometimes include much redundant or useless data, which unnecessarily increases storage space and affects the subsequent data analysis. For example, high redundancy is very common among datasets collected by sensors for environment monitoring. Data compression techniques can be applied to reduce the redundancy. Therefore, data pre-processing operations are indispensable to ensure efficient data storage and exploitation.

3.2.1 Data Collection

Data collection is to utilize special data collection techniques to acquire raw data from a specific data generation environment. Four common data collection methods are shown as follows.

- *Log Files*: As one widely used data collection method, log files are record files automatically generated by the data source system, so as to record activities in designated file formats for subsequent analysis. Log files are typically used in nearly all digital devices. For example, web servers record in log files number of clicks, click rates, visits, and other property records of web users [7]. To capture activities of users at the web sites, web servers mainly include the following three log file formats: public log file format (NCSA), expanded log format (W3C), and IIS log format (Microsoft). All the three types of log files are in the ASCII text format. Databases other than text files may sometimes be used to store log information to improve the query efficiency of the massive log store [8,9]. There are also some other log files based on data collection, including stock indicators in financial applications and determination of operating states in network monitoring and traffic management.
- *Sensors*: Sensors are common in daily life to measure physical quantities and transform physical quantities into readable digital signals for subsequent processing (and storage). Sensory data may be classified as sound wave, voice, vibration, automobile, chemical, current, weather, pressure, temperature, etc. Sensed information is transferred to a data collection point through wired or wireless networks. For applications that may be easily deployed and managed, e.g., video surveillance system [10], the wired sensor network is a convenient solution to acquire related information. Sometimes the accurate position of a specific phenomenon is unknown, and sometimes the monitored environment does not have the energy or communication infrastructures. Then wireless communication must be used to enable data transmission among sensor nodes under limited energy and communication capability. In recent years, WSNs have received considerable interest and have been applied to many applications, such as environmental research [11, 12], water quality monitoring [13], civil engineering [14, 15], and wildlife habit monitoring [16]. A WSN generally consists of a large number of geographically distributed sensor nodes, each being a micro device powered by battery. Such sensors are deployed at designated positions as required by the application to collect remote sensing data. Once the sensors are deployed, the base station will send control information for network configuration/management or data collection to sensor nodes. Based on such control information, the sensory data is assembled in different sensor nodes and sent back to the base station for further processing. Interested readers are referred to [17] for more detailed discussions.
- *Methods for Acquiring Network Data*: At present, network data acquisition is accomplished using a combination of web crawler, word segmentation system, task system, and index system, etc. Web crawler is a program used by search engines for downloading and storing web pages [18]. Generally speaking, web crawler starts from the uniform resource locator (URL) of an initial web page to access other linked web pages, during which it stores and sequences all the retrieved URLs. Web crawler acquires a URL in the order of precedence through a URL queue and then downloads web pages, and identifies all URLs in the downloaded web pages, and extracts new URLs to be put in the queue. This

process is repeated until the web crawler is stopped. Data acquisition through a web crawler is widely applied in applications based on web pages, such as search engines or web caching. Traditional web page extraction technologies feature multiple efficient solutions and considerable research has been done in this field. As more advanced web page applications are emerging, some extraction strategies are proposed in [19] to cope with rich Internet applications.

The current network data acquisition technologies mainly include traditional Libpcap-based packet capture technology, zero-copy packet capture technology, as well as some specialized network monitoring software such as Wireshark, SmartSniff, and WinNetCap.

- *Libpcap-Based Packet Capture Technology*: Libpcap (packet capture library) is a widely used network data packet capture function library. It is a general tool that does not depend on any specific system and is mainly used to capture data in the data link layer. It features simplicity, easy-to-use, and portability, but has a relatively low efficiency. Therefore, under a high-speed network environment, considerable packet losses may occur when Libpcap is used.
- *Zero-Copy Packet Capture Technology*: The so-called zero-copy (ZC) means that no copies between any internal memories occur during packet receiving and sending at a node. In sending, the data packets directly start from the user buffer of applications, pass through the network interfaces, and arrive at an external network. In receiving, the network interfaces directly send data packets to the user buffer. The basic idea of zero-copy is to reduce data copy times, reduce system calls, and reduce CPU load while datagrams are passed from network equipments to user program space. The zero-copy technology first utilizes direct memory access (DMA) technology to directly transmit network datagrams to an address space pre-allocated by the system kernel, so as to avoid the participation of CPU. In the meanwhile, it maps the internal memory of the datagrams in the system kernel to the that of the detection program, or builds a cache region in the user space and maps it to the kernel space. Then the detection program directly accesses the internal memory, so as to reduce internal memory copy from system kernel to user space and reduce the amount of system calls.
- *Mobile Equipments*: At present, mobile devices are more widely used. As mobile device functions become increasingly stronger, they feature more complex and multiple means of data acquisition as well as more variety of data. Mobile devices may acquire geographical location information through positioning systems; acquire audio information through microphones; acquire pictures, videos, streetscapes, two-dimensional barcodes, and other multimedia information through cameras; acquire user gestures and other body language information through touch screens and gravity sensors. Over the years, wireless operators have improved the service level of the mobile Internet by acquiring and analyzing such information. For example, iPhone itself is a "mobile spy." It may collect wireless data and geographical location information, and then send such information back to Apple Inc. for processing, of which the user may not be

aware. Apart from Apple, smart phone operating systems such as Android of Google and Windows Phone of Microsoft can also collect information in the similar manner.

In addition to the aforementioned three data acquisition methods of main data sources, there are many other data collect methods or systems. For example, in scientific experiments, many special tools can be used to collect experimental data, such as magnetic spectrometers and radio telescopes. We may classify data collection methods from different perspectives. From the perspective of data sources, data collection methods can be classified into two categories: collection methods recording through data sources and collection methods recording through other auxiliary tools.

3.2.2 Data Transportation

Upon the completion of raw data collection, data will be transferred to a data storage infrastructure for processing and analysis. As discussed in Sect. 2.3, big data is mainly stored in a data center. The data layout should be adjusted to improve computing efficiency or facilitate hardware maintenance. In other words, internal data transmission may occur in the data center. Therefore, data transmission consists of two phases: Inter-DCN transmissions and Intra-DCN transmissions.

Inter-DCN transmissions are from data source to data center, which is generally achieved with the existing physical network infrastructure. Because of the rapid growth of traffic demands, the physical network infrastructure in most regions around the world are constituted by high-volume, high-rate, and cost-effective optic fiber transmission systems. Over the past 20 years, advanced management equipment and technologies have been developed, such as IP-based wavelength division multiplexing (WDM) network architecture, to conduct smart control and management of optical fiber networks [20, 21]. WDM is a technology that multiplexes multiple optical carrier signals with different wave lengths and couples them to the same optical fiber of the optical link. In such technology, lasers with different wave lengths carry different signals. By far, the backbone network have been deployed with WDM optical transmission systems with single channel rate of 40 Gb/s. At present, 100 Gb/s commercial interface are available and 100 Gb/s systems (or TB/s systems) will be available in the near future [22].

However, traditional optical transmission technologies are limited by the bandwidth of the electronic bottleneck [23]. Recently, orthogonal frequency-division multiplexing (OFDM), initially designed for wireless systems, is regarded as one of the main candidate technologies for future high-speed optical transmission. OFDM is a multi-carrier parallel transmission technology. It segments a high-speed data flow to transform it into low-speed sub-data-flows to be transmitted over multiple orthogonal sub-carriers [24]. Compared with fixed channel spacing of WDM, OFDM allows sub-channel frequency spectrums to overlap with each other [25]. Therefore, it is a flexible and efficient optical networking technology.

Intra-DCN transmissions are the data communication flows within data centers. Intra-DCN transmissions depend on the communication mechanism within the data center (i.e., on physical connection plates, chips, internal memories of data servers, network architectures of data centers, and communication protocols). A data center consists of multiple integrated server racks interconnected with its internal connection networks. Nowadays, the internal connection networks of most data centers are fat-tree, two-layer or three-layer structures based on multi-commodity network flows [23, 26]. In the two-layer topological structure, the racks are connected by 1 Gbps top rack switches (TOR) and then such top rack switches are connected with 10 Gbps aggregation switches in the topological structure. The three-layer topological structure is a structure augmented with one layer on the top of the two-layer topological structure and such layer is constituted by 10 or 100 Gbps core switches to connect aggregation switches in the topological structure. There are also other topological structures which aim to improve the data center networks [27–30].

Because of the inadequacy of electronic packet switches, it is difficult to increase communication bandwidths while keeps energy consumption is low. Over the years, due to the huge success achieved by optical technologies, the optical interconnection among the networks in data centers has drawn great interest. Optical interconnection is a high-throughput, low-delay, and low-energy-consumption solution. At present, optical technologies are only used for point-to-point links in data centers. Such optical links provide connection for the switches using the low-cost multi-mode fiber (MMF) with 10 Gbps data rate. Optical interconnection (switching in the optical domain) of networks in data centers is a feasible solution, which can provide Tbps-level transmission bandwidth with low energy consumption.

Recently, many optical interconnection plans are proposed for data center networks [31]. Some plans add optical paths to upgrade the existing networks, and other plans completely replace the current switches [31–36]. As a strengthening technology, Zhou et al. in [37] adopt wireless links in the 60 GHz frequency band to strengthen wired links. Network virtualization should also be considered to improve the efficiency and utilization of data center networks.

3.2.3 Data Pre-processing

Because of the wide variety of data sources, the collected datasets vary with respect to noise, redundancy, and consistency, etc., and it is undoubtedly a waste to store meaningless data. In addition, some analytical methods have stringent requirements on data quality. Therefore, data should be pre-processed under many circumstances to integrate the data from different sources, so as to enable effective data analysis. Pre-processing data not only reduces storage expense, but also improves analysis accuracy. Some relational data pre-processing techniques are discussed in the following.

3.2.3.1 Integration

Data integration is the cornerstone of modern commercial informatics, which involves the combination of data from different sources and provides users with a uniform view of data [38]. This is a mature research field for traditional database. Historically, two methods have been widely recognized: data warehouse and data federation. Data warehousing includes a process named ETL (Extract, Transform and Load). Extraction involves connecting source systems, selecting, collecting, analyzing, and processing necessary data. Transformation is the execution of a series of rules to transform the extracted data into standard formats. Loading means importing extracted and transformed data into the target storage infrastructure. Loading is the most complex procedure among the three, which includes operations such as transformation, copy, clearing, standardization, screening, and data organization. A virtual database can be built to query and aggregate data from different data sources, but such database does not contain data. On the contrary, it includes information or metadata related to actual data and its positions. Such two "storage-reading" approaches do not satisfy the high performance requirements of data flows or search programs and applications. Compared with queries, data in such two approaches is more dynamic and must be processed during data transmission. Generally, data integration methods are accompanied with flow processing engines and search engines [39, 40].

3.2.3.2 Cleaning

Data cleaning is a process to identify inaccurate, incomplete, or unreasonable data, and then modify or delete such data to improve data quality. Generally, data cleaning includes five complementary procedures [41]: defining and determining error types, searching and identifying errors, correcting errors, documenting error examples and error types, and modifying data entry procedures to reduce future errors. During cleaning, data formats, completeness, rationality, and restriction shall be inspected. Data cleaning is of vital importance to keep the data consistency, which is widely applied in many fields, such as banking, insurance, retail industry, telecommunications, and traffic control.

In e-commerce, most data is electronically collected, which may have serious data quality problems. Classic data quality problems mainly come from software defects, customized errors, or system mis-configuration. Authors in [42] discussed data cleaning in e-commerce by crawlers and regularly re-copying customer and account information. In [43], the problem of cleaning RFID data was examined. RFID is widely used in many applications, e.g., inventory management and target tracking. However, the original RFID features low quality, which includes a lot of abnormal data limited by the physical design and affected by environmental noises. In [44], a probability model was developed to cope with data loss in mobile environments. Khoussainova et al. in [45] proposed a system to automatically correct errors of input data by defining global integrity constraints. Herbert et al. [46]

proposed a framework called BIO-AJAX to standardize biological data so as to conduct further computation and improve search quality. With BIO-AJAX, some errors and repetitions may be eliminated, and common data mining technologies can be executed more effectively.

3.2.3.3 Redundancy Elimination

Data redundancy refers to data repetitions or surplus, which usually occurs in many datasets. Data redundancy can increase the unnecessary data transmission expense and cause defects on storage systems, e.g., waste of storage space, leading to data inconsistency, reduction of data reliability, and data damage. Therefore, various redundancy reduction methods have been proposed, such as redundancy detection, data filtering, and data compression. Such methods may apply to different datasets or application environments. However, redundancy reduction may also bring about certain negative effects. For example, data compression and decompression cause additional computational burden. Therefore, the benefits of redundancy reduction and the cost should be carefully balanced.

Data collected from different fields will increasingly appear in image or video formats. It is well-known that images and videos contain considerable redundancy, including temporal redundancy, spacial redundancy, statistical redundancy, and sensing redundancy. Video compression is widely used to reduce redundancy in video data, as specified in the many video coding standards (MPEG-2, MPEG-4, H.263, and H.264/AVC). In [47], the authors investigated the problem of video compression in a video surveillance system with a video sensor network. The authors propose a new MPEG-4 based method by investigating the contextual redundancy related to background and foreground in a scene. The low complexity and the low compression ratio of the proposed approach were demonstrated by the evaluation results.

On generalized data transmission or storage, repeated data deletion is a special data compression technology, which aims to eliminate repeated data copies [48]. With repeated data deletion, individual data blocks or data segments will be assigned with identifiers (e.g., using a hash algorithm) and stored, with the identifiers added to the identification list. As the analysis of repeated data deletion continues, if a new data block has an identifier that is identical to that listed in the identification list, the new data block will be deemed as redundant and will be replaced by the corresponding stored data block. Repeated data deletion can greatly reduce storage requirement, which is particularly important to a big data storage system. Apart from the aforementioned data pre-processing methods, specific data objects shall go through some other operations such as feature extraction. Such operation plays an important role in multimedia search and DNA analysis [49–51]. Usually high-dimensional feature vectors (or high-dimensional feature points) are used to describe such data objects and the system stores the dimensional feature vectors for future retrieval. Data transfer is usually used to process distributed heterogeneous data sources, especially business datasets [52].

As a matter of fact, in consideration of various datasets, it is non-trivial, or impossible, to build a uniform data pre-processing procedure and technology that is applicable to all types of datasets. on the specific feature, problem, performance requirements, and other factors of the datasets should be considered, so as to select a proper data pre-processing strategy.

References

1. James Manyika, McKinsey Global Institute, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.
2. John Gantz and David Reinsel. The digital universe decade-are you ready. *External Publication of IDC (Analyse the Future) Information and Data*, pages 1–16, 2010.
3. Douglas Laney. 3-d data management: Controlling data volume, velocity and variety. *META Group Research Note, February*, 6, 2001.
4. Kenneth Cukier. *Data, data everywhere: A special report on managing information*. Economist Newspaper, 2010.
5. Paul Zikopoulos, Chris Eaton, et al. *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media, 2011.
6. Randal E Bryant. Data-intensive scalable computing for scientific applications. *Computing in Science & Engineering*, 13(6):25–33, 2011.
7. Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, and Mohamad Farhan Mohamad Mohsin. Data pre-processing on web server logs for generalized association rules mining algorithm. *World Academy of Science, Engineering and Technology*, 48:2008, 2008.
8. Alexandros Nanopoulos, Yannis Manolopoulos, Maciej Zakrzewicz, and Tadeusz Morzy. Indexing web access-logs for pattern queries. In *Proceedings of the 4th international workshop on Web information and data management*, pages 63–68. ACM, 2002.
9. Karuna P Joshi, Anupam Joshi, and Yelena Yesha. On using a warehouse to analyze web logs. *Distributed and Parallel Databases*, 13(2):161–180, 2003.
10. Vijay Chandramohan and Ken Christensen. A first look at wired sensor networks for video surveillance systems. In *Local Computer Networks, 2002. Proceedings. LCN 2002. 27th Annual IEEE Conference on*, pages 728–729. IEEE, 2002.
11. Leo Selavo, Anthony Wood, Qing Cao, Tamim Sookoor, Hengchang Liu, Aravind Srinivasan, Yafeng Wu, Woochul Kang, John Stankovic, Don Young, et al. Luster: wireless sensor network for environmental research. In *Proceedings of the 5th international conference on Embedded networked sensor systems*, pages 103–116. ACM, 2007.
12. Guillermo Barrenetxea, François Ingelrest, Gunnar Schaefer, Martin Vetterli, Olivier Couach, and Marc Parlange. Sensorscope: Out-of-the-box environmental monitoring. In *Information Processing in Sensor Networks, 2008. IPSN'08. International Conference on*, pages 332–343. IEEE, 2008.
13. Younghun Kim, Thomas Schmid, Zainul M Charbiwala, Jonathan Friedman, and Mani B Srivastava. Nawms: nonintrusive autonomous water monitoring system. In *Proceedings of the 6th ACM conference on Embedded network sensor systems*, pages 309–322. ACM, 2008.
14. Sukun Kim, Shamim Pakzad, David Culler, James Demmel, Gregory Fenves, Steven Glaser, and Martin Turon. Health monitoring of civil infrastructures using wireless sensor networks. In *Information Processing in Sensor Networks, 2007. IPSN 2007. 6th International Symposium on*, pages 254–263. IEEE, 2007.
15. Matteo Ceriotti, Luca Mottola, Gian Pietro Picco, Amy L Murphy, Stefan Guna, Michele Corra, Matteo Pozzi, Daniele Zonta, and Paolo Zanon. Monitoring heritage buildings with wireless

sensor networks: The torre aquila deployment. In *Proceedings of the 2009 International Conference on Information Processing in Sensor Networks*, pages 277–288. IEEE Computer Society, 2009.
16. Gilman Tolle, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, Todd Dawson, Phil Buonadonna, David Gay, et al. A macroscope in the redwoods. In *Proceedings of the 3rd international conference on Embedded networked sensor systems*, pages 51–63. ACM, 2005.
17. Feng Wang and Jiangchuan Liu. Networked wireless sensor data collection: issues, challenges, and approaches. *Communications Surveys & Tutorials, IEEE*, 13(4):673–687, 2011.
18. Junghoo Cho and Hector Garcia-Molina. Parallel crawlers. In *Proceedings of the 11th international conference on World Wide Web*, pages 124–135. ACM, 2002.
19. Suryakant Choudhary, Mustafa Emre Dincturk, Seyed M Mirtaheri, Ali Moosavi, Gregor von Bochmann, Guy-Vincent Jourdan, and Iosif-Viorel Onut. Crawling rich internet applications: the state of the art. In *CASCON*, pages 146–160, 2012.
20. Nasir Ghani, Sudhir Dixit, and Ti-Shiang Wang. On ip-over-wdm integration. *Communications Magazine, IEEE*, 38(3):72–84, 2000.
21. James Manchester, Jon Anderson, Bharat Doshi, and Subra Dravida. Ip over sonet. *Communications Magazine, IEEE*, 36(5):136–142, 1998.
22. M Jinno, H Takara, and B Kozicki. Dynamic optical mesh networks: Drivers, challenges and solutions for the future. In *Optical Communication, 2009. ECOC'09. 35th European Conference on*, pages 1–4. IEEE, 2009.
23. Luiz André Barroso and Urs Hölzle. The datacenter as a computer: An introduction to the design of warehouse-scale machines. *Synthesis Lectures on Computer Architecture*, 4(1):1–108, 2009.
24. Jean Armstrong. Ofdm for optical communications. *Journal of lightwave technology*, 27(3): 189–204, 2009.
25. William Shieh. Ofdm for flexible high-speed optical networks. *Journal of Lightwave Technology*, 29(10):1560–1577, 2011.
26. Cisco data center interconnect design and deployment guide, 2010.
27. Albert Greenberg, James R Hamilton, Navendu Jain, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, David A Maltz, Parveen Patel, and Sudipta Sengupta. Vl2: a scalable and flexible data center network. In *ACM SIGCOMM Computer Communication Review*, volume 39, pages 51–62. ACM, 2009.
28. Chuanxiong Guo, Guohan Lu, Dan Li, Haitao Wu, Xuan Zhang, Yunfeng Shi, Chen Tian, Yongguang Zhang, and Songwu Lu. Bcube: a high performance, server-centric network architecture for modular data centers. *ACM SIGCOMM Computer Communication Review*, 39(4):63–74, 2009.
29. Nathan Farrington, George Porter, Sivasankar Radhakrishnan, Hamid Hajabdolali Bazzaz, Vikram Subramanya, Yeshaiahu Fainman, George Papen, and Amin Vahdat. Helios: a hybrid electrical/optical switch architecture for modular data centers. *ACM SIGCOMM Computer Communication Review*, 41(4):339–350, 2011.
30. Hussam Abu-Libdeh, Paolo Costa, Antony Rowstron, Greg O'Shea, and Austin Donnelly. Symbiotic routing in future data centers. *ACM SIGCOMM Computer Communication Review*, 40(4):51–62, 2010.
31. Cedric Lam, Hong Liu, Bikash Koley, Xiaoxue Zhao, Valey Kamalov, and Vijay Gill. Fiber optic communication technologies: What's needed for datacenter network operations. *Communications Magazine, IEEE*, 48(7):32–39, 2010.
32. Guohui Wang, David G Andersen, Michael Kaminsky, Konstantina Papagiannaki, TS Ng, Michael Kozuch, and Michael Ryan. c-through: Part-time optics in data centers. In *ACM SIGCOMM Computer Communication Review*, volume 40, pages 327–338. ACM, 2010.
33. Xiaohui Ye, Yawei Yin, SJ Ben Yoo, Paul Mejia, Roberto Proietti, and Venkatesh Akella. Dos: A scalable optical switch for datacenters. In *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems*, page 24. ACM, 2010.

34. Ankit Singla, Atul Singh, Kishore Ramachandran, Lei Xu, and Yueping Zhang. Proteus: a topology malleable data center network. In *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, page 8. ACM, 2010.
35. Odile Liboiron-Ladouceur, Isabella Cerutti, Pier Giorgio Raponi, Nicola Andriolli, and Piero Castoldi. Energy-efficient design of a scalable optical multiplane interconnection architecture. *Selected Topics in Quantum Electronics, IEEE Journal of*, 17(2):377–383, 2011.
36. Avinash Karanth Kodi and Ahmed Louri. Energy-efficient and bandwidth-reconfigurable photonic networks for high-performance computing (hpc) systems. *Selected Topics in Quantum Electronics, IEEE Journal of*, 17(2):384–395, 2011.
37. Xia Zhou, Zengbin Zhang, Yibo Zhu, Yubo Li, Saipriya Kumar, Amin Vahdat, Ben Y Zhao, and Haitao Zheng. Mirror mirror on the ceiling: Flexible wireless links for data centers. *ACM SIGCOMM Computer Communication Review*, 42(4):443–454, 2012.
38. Maurizio Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM, 2002.
39. Wiki. Applications and organizations using hadoop. http://wiki.apache.org/hadoop/PoweredBy, 2013.
40. Michael J Cafarella, Alon Halevy, and Nodira Khoussainova. Data integration for the relational web. *Proceedings of the VLDB Endowment*, 2(1):1090–1101, 2009.
41. Jonathan I Maletic and Andrian Marcus. Data cleansing: Beyond integrity analysis. In *IQ*, pages 200–209. Citeseer, 2000.
42. Ron Kohavi, Llew Mason, Rajesh Parekh, and Zijian Zheng. Lessons and challenges from mining retail e-commerce data. *Machine Learning*, 57(1–2):83–113, 2004.
43. Haiquan Chen, Wei-Shinn Ku, Haixun Wang, and Min-Te Sun. Leveraging spatio-temporal redundancy for rfid data cleansing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 51–62. ACM, 2010.
44. Zhou Zhao and Wilfred Ng. A model-based approach for rfid data stream cleansing. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 862–871. ACM, 2012.
45. Nodira Khoussainova, Magdalena Balazinska, and Dan Suciu. Probabilistic event extraction from rfid data. In *Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on*, pages 1480–1482. IEEE, 2008.
46. Katherine G Herbert and Jason TL Wang. Biological data cleaning: a case study. *International Journal of Information Quality*, 1(1):60–82, 2007.
47. Tsung-Han Tsai and Chung-Yuan Lin. Exploring contextual redundancy in improving object-based video coding for video sensor networks surveillance. *Multimedia, IEEE Transactions on*, 14(3):669–682, 2012.
48. Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278. ACM, 2002.
49. Uday Kamath, Jack Compton, Rezarta Islamaj Dogan, Kenneth De Jong, and Amarda Shehu. An evolutionary algorithm approach for feature generation from sequence data and its application to dna splice site prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(5):1387–1398, 2012.
50. Kwong-Sak Leung, Kin Hong Lee, Jin-Feng Wang, Eddie YT Ng, Henry LY Chan, Stephen KW Tsui, Tony SK Mok, PC-H Tse, and JJ-Y Sung. Data mining on dna sequences of hepatitis b virus. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(2):428–440, 2011.
51. Zi Huang, Hengtao Shen, Jiajun Liu, and Xiaofang Zhou. Effective data co-reduction for multimedia similarity search. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 1021–1032. ACM, 2011.
52. Jens Bleiholder and Felix Naumann. Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1, 2008.