Survey of Recent Research Progress and Issues in Big Data

Bo Li, boli@seas.wustl.edu (A paper written under the guidance of Prof. Raj Jain)

Download

Abstract

Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications. This paper reveals most recent progress on big data networking and big data. We have categorized reported efforts into four general categories. First, efforts related to classic big data technology such as storage, Software-Defined Network, data transportation and analytics are reported. Second, important aspects of big data in cloud computing such as recourse management and performances optimization are introduced. Lastly, we introduce interesting benchmarks and progress in both search engines and mobile networking. Upon detailed summary and analysis, limitations of the proposed works as well as possible future research directions have been proposed.

Key Words

Big Data, Big Data Networking, Hadoop, MapReduce, Cloud Computing, Benchmark, Mobile Networking

Table of Contents

Lists of Figures and Tables

Figure 3. System Architecture of Radoop

Figure 4. Multi-tenant Model in Cloud Computing

1. Introduction

Big data is the term for data sets so large and complicated that it becomes difficult to process using traditional data management tools or processing applications. This paper reveals recent progress on big data, big data networking and relevant topics.

According to [Bakshi12], the size of digital data in 2011 is roughly 1.8 Zettabytes (1.8 trillion gigabytes). That is, supporting networking infrastructure has to manage 50 times more information by year 2020. Specifically, considerations of efficiency, economics and privacy should be carefully planned while including new big data building blocks into existing data and networking infrastructure [Bakshi12].

In addition to big data challenges induced by traditional data generation, consumption, and analytics at a much larger scale, newly emerged characteristics of big data has shown important trends on mobility of data, faster data access and consumption, as well as ecosystem capabilities [Cisco11]. Fig. 1 illustrates a general big data network model with MapReduce. Distinct applications in the cloud has put demanding requirements for acquisition, transportation and analytics of structured and unstructured data.

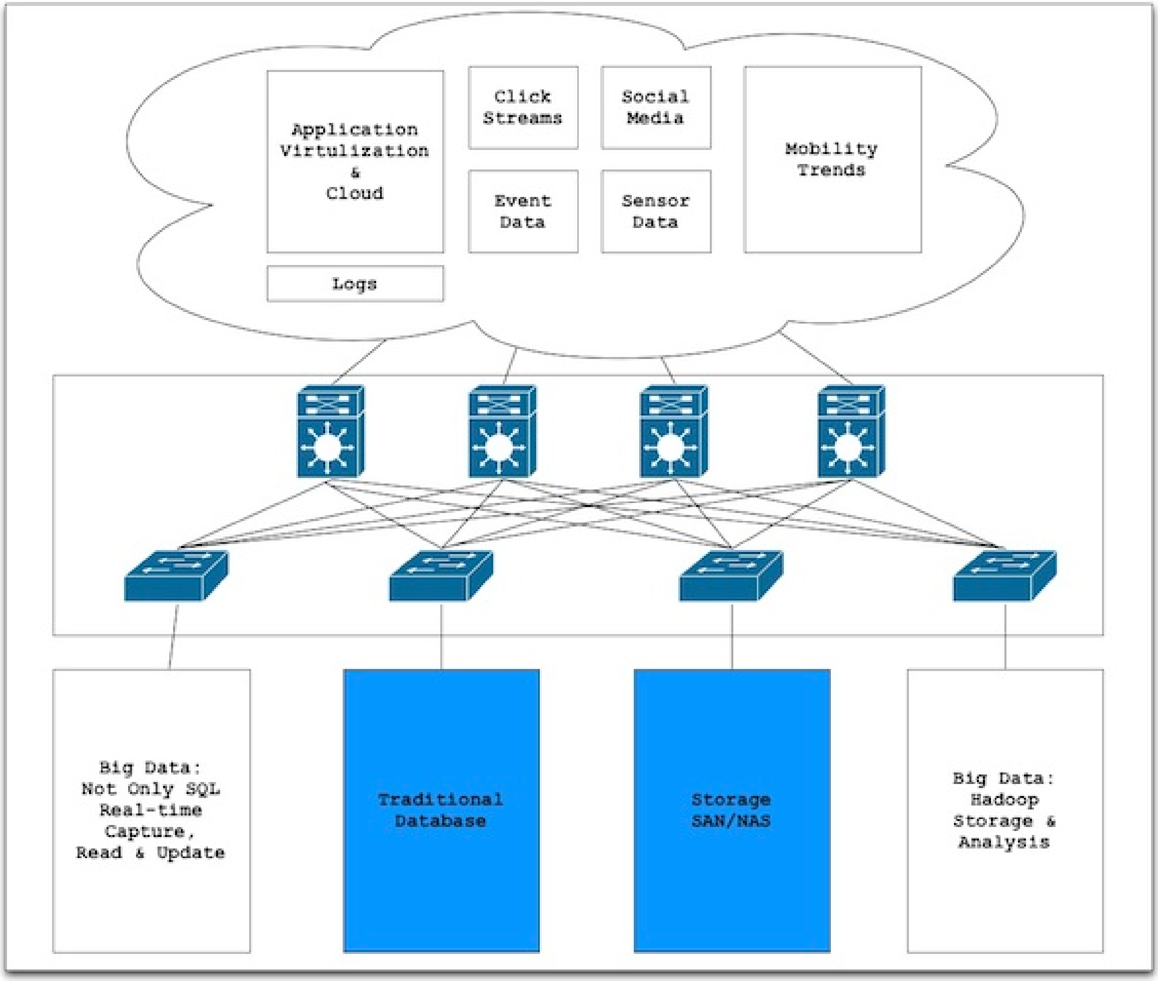
Application Virtulization & Cloud
Click Streams
Social Media
Event Data
Sensor Data
Mobility Trends
Logs
Big Data: Not Only SQL Real-time Capture, Read & Update
Traditional Database
Storage SAN/NAS
Big Data: Hadoop Storage & Analysis

Figure 1: General Framework of Big Data Networking.

In this paper, we pay close attention to recent progresses made on big data and big data networking. We divide relevant efforts into representative categories while maintaining our own independent understandings. To be specific, topics covered in this paper include: recent progress on classic big data networking technologies, e.g., Hadoop and MapReduce, big data technologies in could computing, big data benchmarking projects, and mobile big data networking.

2. Related Work

This section reveals recent progress and efforts in big data networking. We cover these topics in 4 categories: classic big data networking technology, big data in cloud computing, data engineering and benchmarking approaches, and mobile big data networking. All covered topics are reported between 2011 and 2013.

As classical big data research, the following work reported progress in big data networking. [Madden12] reveals challenges and opportunities in databases in existence of big data. [Girola11] introduces virtualization planning and cloud computing methods in IBM data center networking. [Keim13] depicts interesting methodologies in big data virtualization. From a platform architecting perspective, [Ferguson12] reports their progress for accelerating big data analytics. As a recent effort, [Dittrich12] introduces contributions on optimizing big data processing efficiency in Hadoop and MapReduce.

There have been a number of new and interesting big data methodologies reported. [Monga12] introduces their efforts on Software-Defined Networking for big-data science-architectural models in campus environment and more importantly in Wide Area Network. [Herodotou11] proposed a self-tuning system for big data analytics. RCFile as a fast and space-efficient data placement structure in MapReduce warehouse has been proposed in [He11]. An efficient in-network aggregation method for big data applications was introduced in [Costa12], which considerably reduce sizes of data transportation. [Brunet12] reported their method of Gaia Hadoop solution with an emphasis on identifying potential challenges. An interesting application of using big-data for kinect training was discussed in [Budiu12]. [Wang12] introduces their efforts of run-time networking programming in big data applications. A recent case study for bursting data in Transportation SDN was introduced in [Sadasivarao13]. Efforts on optimizing interactions with big data analytics were reported in [Fisher12]. [Begoli12] presented their design principles for efficient knowledge discovery. Radoop, based on RapidMiner and Hadoop, has attracted attention in data analytics [Prekopcsák11]. General considerations for big data architecture and data management has been reported in [Bakshi12]

Remarkable progress of big data networking has also been reported in the area of cloud computing. [Agrawal11] reported existing states and potential future opportunities for big data and cloud computing. Resource management and allocation in multi-cluster clouds were introduced in [Lakew13]. A dataflow-based performance analysis for big data cloud, i.e., Hitune, was presented in [Dai11]. Interesting case studies on big data processing in cloud computing environment was depicted in [Ji12]. [Lu11] presented their work of a framework for cloud-based large-scale data analytics and visualization; a case study on climate data of various scales were introduced too. A recent online cost-minimization approach was depicted in [Zhang13]. Specifically for reducing cooling energy cost for big data analytics cloud, a data-central approach was introduced in [Kaushik12].

In addition to methodologies, there have been a few interesting data engineering and benchmarking efforts reported for big data. [Gao13] introduces a big data benchmark project based on open-source data interfaces of web search engines. [Laurila12] presents a mobile data collection challenge initiated by Nokia, which represents an important step towards mobile big data networking.

Given that mobile networking is becoming a more and more important counterpart of traditional Internet and big data. Big data benchmarking has valuable impact for the research community. [Shekhar12] reported a spatial big-data challenges intersecting mobility and cloud computing. A recent effort on mining large-scale smartphone and data for personality studies has been presented in [Chittaranjan13]. From a perspective of big data applications, [Silberstein11] introduced challenges in social applications while [Zaslavsky13] presented an interesting application as a service of big data.

3. Efforts in Classic Big Data Networking

In addition to traditional big data technologies such as Hadoop, MapReduce and NoSQL, plausible progresses have been made in the past two years on big data networking in many other areas. We summarize them into 4 categories: storage and warehouse, data transportation, Software-Defined Networking and big data Analytics.

3.1 Storage and Warehouse

Data storage is the basis for big data networking. Representative technologies are Relational database and Not Only SQL (NoSQL) databases and data warehouse.

An in-depth review on state-of-art database technologies in the area of big data was presented in [Madden12]. The author claimed that although considerable progresses have been made in database research, much remains to be done: firstly, handling streaming high-rate data in relational models remains as an open problem; second, statistical analysis and machine learning algorithms for big data need to be more robust and easier to use; lastly but more importantly, an ecosystem-alike mechanism should be built around the devised big data algorithms such that data management and usage can evolve sitting on top of the proposed algorithms.

Another important aspect in big data related database is data placement structures. Authors in [He11] argues that traditional data placement structures such as row-stores, column-stores and hybrid-stores are no longer suitable in large data analysis using MapReduce on distributed systems. Instead, the authors have proposed RCFile (Record Columnar File) and its implementation in Hadoop, which meets fast data loading, query processing, efficient storage space utilization, and strong adaptability to dynamic workload patterns[He11]. Basic idea of RCFile is depicted as in Fig. 2.

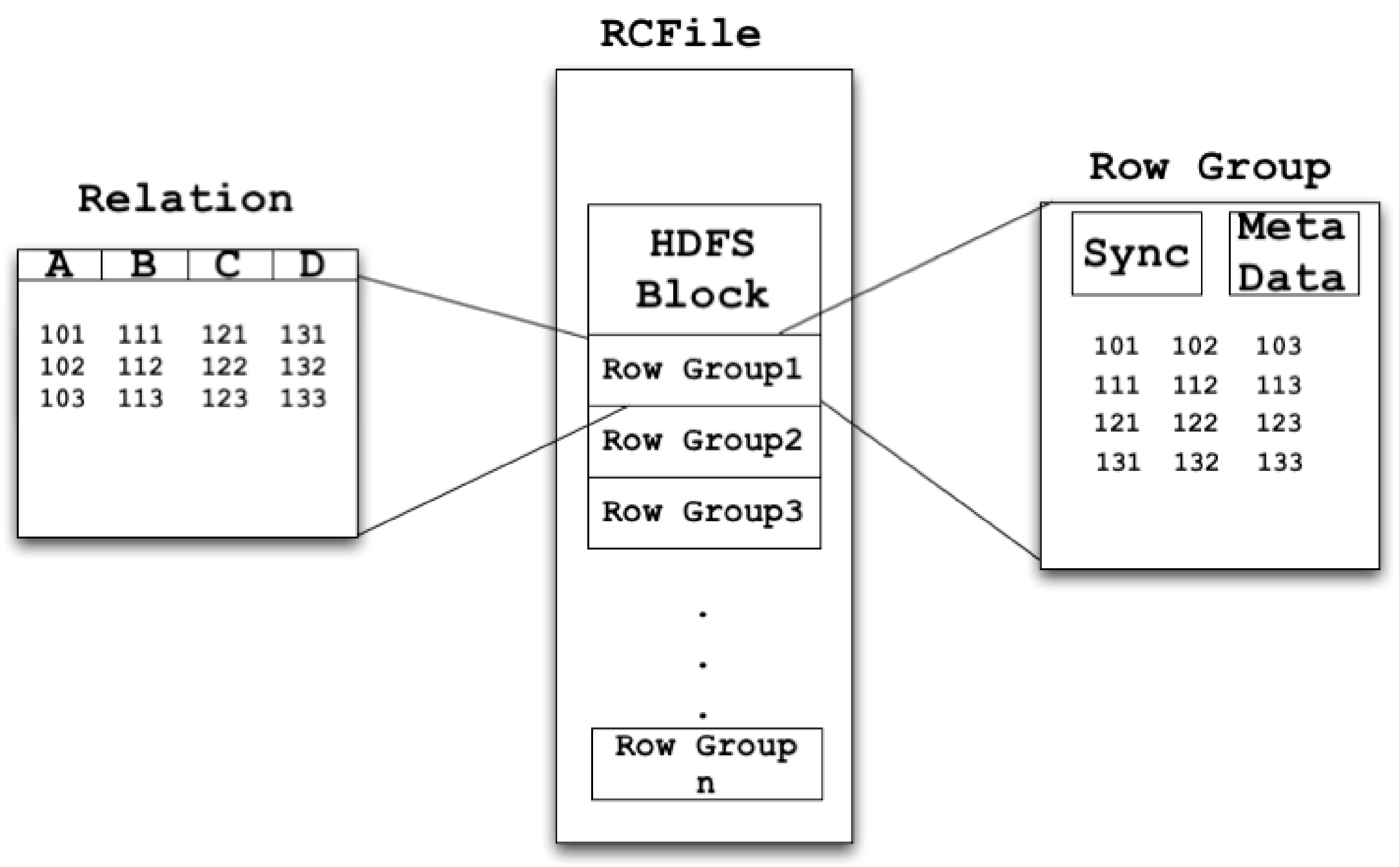

Figure 2. Table Architecture of RCFile.

As in Fig. 2, tables in HDFS of RCFile have multiple HDFS blocks, and each HDFS block is organized with basic units of row groups and all groups have the same size. This clustering idea enables RCFile to more efficiently manage data rows. As in Fig. 3, data areas of RCFile tables are divided as sync marker, metadata and table data sections. More importantly, RCFile has adopted RLE (Run Length Encoding) algorithm to compress metadata while using the Gzip compression algorithm for independently column data compression, which takes advantage the columnar storage of data. Moreover, because of the lazy decompression, RCFile does not need to decompress all columns while processing a row group. Decompression overhead can thus be reduced.

RCFile has been selected as the default data placement method in Facebook. It has also been adopted by Hive and Pig. However, RCFile can still be optimized. For example, currently RCFile does not support arbitrary writings since HDFS currently supports only data writes to the end of files. Automatic selection of the best compression algorithm for each column would be another direction that RCFile can pursue.

Gaia Hadoop in [Brunet12] has included their batch execution framework with HDFS optimized for task execution. Gaia aims to deal with one Petabyte data in tables consisting eighty billion rows. Although the proposed software and hardware design have been proven to work, heterogeneity among hardware remains as a concern and careful design in this regard should be done.

In sum, RCFile and Gaia Hadoop represent recently progress in data storage and warehouse. It is a pleasure to see that fundamental technologies such as data placement and batch data processing can be effectively handled by using appropriate algorithms.

3.2 Software-Defined Network

Software-Defined Network (SDN) as the critical transportation media of big data also plays a critical role in big data networking. We next reveals progress in this regard.

[Monga12] introduces their efforts on Software-Defined Networking for big-data science-architectural models from campus to WAN. To bypass traditional performance hotspots in typical campus network, the authors have built based on the SC11 SCinet Research Sandbox demonstrator with SDN for sake of a scalable architectural approach. The proposed work has been proved to be simple and more importantly adaptable to network framework. Overall speaking, method in this work is incremental, but we are glad to see its system validation has proved yet another SDN design.

Run-time networking programming is useful for big data networks that require frequent reconfigurations. [Wang12] introduces their efforts of run-time networking programming in big data applications. Specifically, the authors combined SDN controller and optical switching to realize close collaboration of network control and potential applications. Joint optimizations of network performance as well as network utilization have been explored. Analysis shows that, at a relatively small overhead of configurations, the proposed integration offers great potentials for optimizing applications performances. The systematic design and evaluation in this work is inspiring.

Bursting data transportation is yet another important aspect for SDN data exchanging as it promises smaller transportation delays. A recent case study for bursting data in Transportation SDN was introduced in [Sadasivarao13]. The authors proposed a SDN-enabled optical transportation architecture which meshes seamlessly within data centers. A case study with an OpenFlow-enabled optical vSwitch managing a small optical transport network was reported. The authors argue that their extension and the inherent programmability brought by SDN are substantial in real world applications. However, general impact has to be further validated in larger deployments.

In sum, real-world case studies on SDN as well as run-time programming and bursting data transportation has been reported and they all showed promising advancement compared to existing approaches. SDN is benefiting from these advancements.

An efficient in-network aggregation method, Camdoop, for big data applications was introduced in [Costa12], which considerably reduce volumes of data transportation. Instead of increasing network bandwidth, authors in this work focused on decreasing the traffic by pushing aggregation from the edge into the network. Implementation based on CamCube and direct-connect topology (i.e., servers connected directly to other servers), Camdoop specifically utilize the property that CamCube servers forward traffics to do in-network aggregation. Case studies showed that Camdoop significantly reduces the network traffic while maintaining comparable performances as opposed to a reference of Hadoop and Dryad/DryadLINQ.

However, similar to tradeoffs in other in-networking aggregation approaches, Camdoop also suffers from losing end-data accuracies, because it does not transport all the generated data. Moreover, in-depth comparisons against more advanced approaches instead of the reported one reference is needed.

3.3 Analytics

Collection and transportation of big data share a common goal: analyzing the data for insights and better application guidance. We reveal new progress as below for big data analytics.

As a recent effort, [Dittrich12] gives a tutorial on optimizing big data processing efficiency in Hadoop and MapReduce. To be specific, the users focused on introducing different data management techniques, e.g., job optimization, physical data organization such as data layouts and indexes. A comprehensive comparison between Hadoop MadReduce and Parallel DBMS was given. From an architecture perspective, [Ferguson12] reported their progress for accelerating big data analytics. This work introduces efforts of IBM in architecting their big data platforms to meet the requirement that one new analytical ecosystem can support entire spectrum of big data analytics. The reported technology utilized Hadoop, IBM Smart Analytic System with built-in NoSQL graph store.

Starfish in [Herodotou11] proposed a self-tuning system for big data analytics. The focus of this work is to mitigate the knowledge gap between new users and the sophisticated configurations of Hadoop and its default MapReduce layer. Moreover, Starfish can adapt to user ends and system workloads for better performance. The basis of Starfish is self-tunning database. Nevertheless, it is not clear how well Starfish can react to high-rate streaming data.

Radoop, based on Rapidminder and Hadoop, has attracted attentions in data analytics [Prekopcsák11]. Integration, development and runtime measurement on a few data transformation tasks have validated feasibility of Radoop for big data analytics with scalable network size and data volumes. System architecture of Radoop can be seen in Fig. 3.

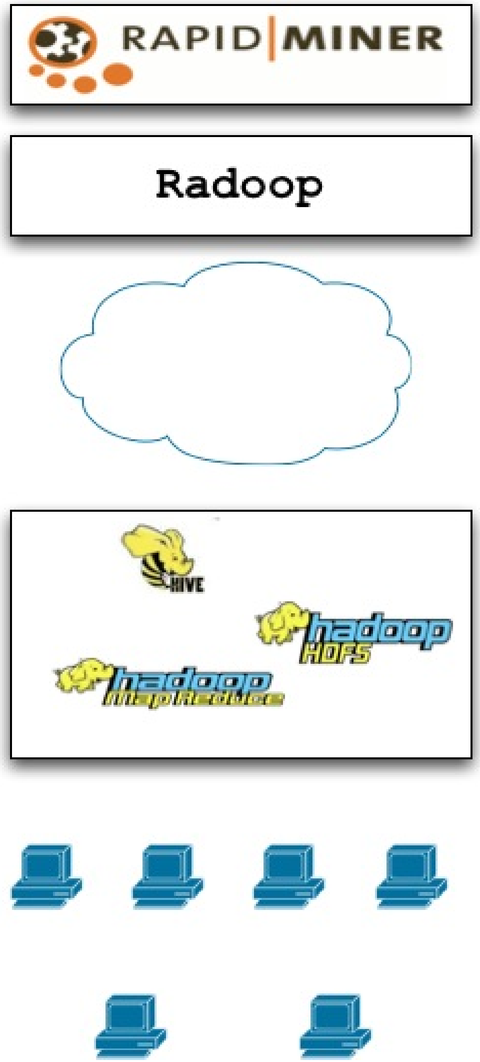

Figure 3. System Architecture of Radoop

As in Fig. 3, integration of RapidMiner and Hadoop has enabled Radoop to fully take advantage of both sides; however, a further step of componentizing Radoop blocks to further leverage cross-layer tradeoff seems to be a promising step.

Specifically for big data analytics, IBM Smart Analytic System, Starfish and Radoop represent one more step towards efficient data management, system adaptation/tunning, and large-scale big data transportation and analysis, respectively. These efforts will enhance the basis of big data and big data networking.

4. Progress of Big Data in Could Computing

Cloud Computing as an important application environment for big data has attracted tremendous attentions from the research community. Remarkable progress of big data networking has also been reported in this area. In this section, we introduce big data research issues and solutions related to Cloud Computing. Specifically, we are interested in the following topics: opportunities and challenges of big data networking in Cloud Computing, cloud resource management of big data, and performance optimization of big data in Cloud Computing.

4.1 Overview and Resource Management

Authors in [Agrawal11] has comprehensively reported existing states and potential future opportunities for big data and cloud computing. Specifically, Agrawal et al. focused on systems for supporting update heavy applications and ad-hoc analytics and decision support. Multi-tenant system model with different level of resource sharing is shown in Fig. 4.

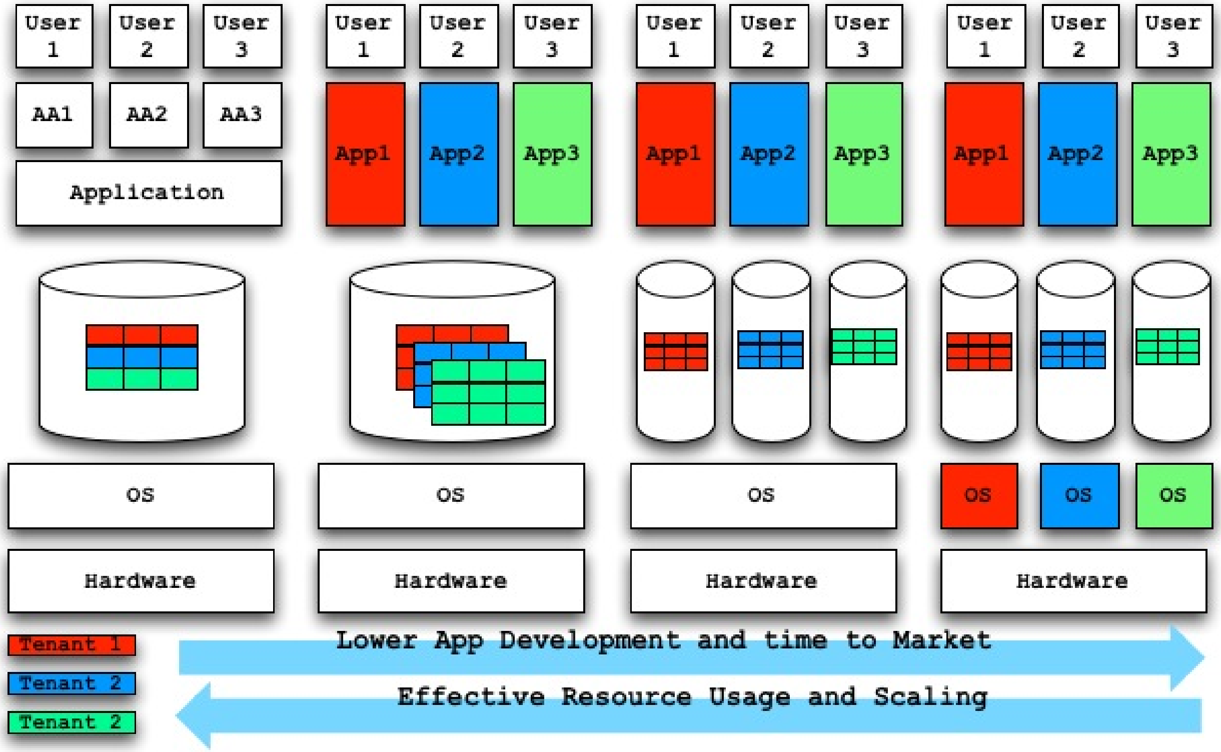

Figure 4. Multi-tenant model: left to right, shared table, shared database, shared OS and shared hardware.

Fig.4 depicts representative forms of the challenging multi-tenant model and trade-offs associated with different forms of sharing. Since models share resources at different levels of abstraction, isolation guarantees can be achieved differently accordingly.

Resource management plays a fundamental role in big data applications in the cloud. We next review important progress in this regard.

A general introduction to resource management and allocation in multi-cluster clouds were introduced in [Lakew13Lakew13]. [Girola11] introduces virtualization planning and cloud computing methods in IBM data center networking. Key operational challenges such as support cost-saving technologies, rapid deployment, support for mobile and pervasive access, development of enterprise-grade network design has been discussed extensively. Despite existing efforts taking care of these challenges, an open question remains for making these objectives possible in a real-time and scalable fashion.

[Lu11] presented their work of a framework for cloud-based large-scale data analytics and virtualization; a case study on climate data of various scales were introduced too. Representative problems such as large datasets versus limited computational resources, data complexity versus limited knowledge, varying data structures/formats versus the need to integrate different tools. A case study with spatial temporal data set validated effectiveness of the proposed framework. Key idea in this work is to make better use of computational and storage resources with the help of componentized software and cross-layer communications, which is as expected.

Specifically for reducing cooling energy cost for big data analytics cloud, a data-centric approach was introduced in [Kaushik12]. Instead of relying on thermal-aware computational job placement/migration, the method in [Kaushik12] takes a data-centric approach, which is now popular in big data applications. While reducing cooling energy costs, thermal reliability servers can been achieved. Evaluation with real big data analytics traces from Yahoo shows 9x better performances than state-of-art data-agnostic cooling methods.

In sum, pervasive computing of big data in the cloud, computational resource and data complexity management, and energy consumption manipulations for big data in the cloud are fundamentally important aspects. The studied works have made plausible progress in terms of system design and implementation, but much remains to be done with consideration of system validation in larger, real-world applications.

4.2 Performance Optimization

Performance optimization is yet another classic and important topic in cloud computing because appropriate optimization techniques will provide better application experiences with comparable or even less system resource consumption, compared to non-optimized cases.

A dataflow-based performance analysis tool for big data cloud, i.e., Hitune, was presented in [Dai11]. Hitune is shown to be effective in assisting users doing Hadoop performance analysis and system parameter tuning. Limitations of existing approaches, such as Hadoop logs and metrics was also compared and discussed. A few interesting case studies on big data processing in cloud computing environment was depicted in [Ji12]. Efforts of the Fijitsu laboratory are based on data store and complex event processing, as well as workflow description in distributed data processing.

A recent online cost-minimization algorithm was depicted in [Zhang13]. The proposed work specifically focused on real-time cost minimizations for uploading massive and dynamic data onto the cloud. The two online algorithms have achieved competitive cost reduction ratios. However, the proposed methods are only evaluated in a limited scale. The proposed algorithms need to be further evaluated at larger and more competitive scales, e.g., data streaming applications with larger topologies.

In sum, Hitune and the Fijitsu laboratory approaches have been focused on promoting user experiences by using fundamental big data techniques such as event processing and work flow description. Tools and case studies like this are informational and offer more choices to users. Moreover, online cost-minimizing as another promising direction has been proved to be effective in big data applications. We expect a lot more scalable and efficient algorithms to be proposed in the near future.

5. Big Data Benchmarks and Mobile Networking

In this section, we briefly reveal two important counterparts of big data networking research: benchmarks and mobile networking with big data considerations. The discussed works represents not only the dedicated efforts but also possible popular trends in big data networking research.

5.1 Big Data Benchmarks

Big data benchmarks play a crucial role in these data-centric research areas, because scientifically collection and organization of informational data will provide important ground truth for further methodology verifications.

Authors in [Gao13] present BigdataBench, an interesting big data benchmark project based on open-source data interfaces of web search engines. As we all know, search engines have been entrance point of the whole Internet. Hence, insightful collection of informational data sets is not only valuable but also hard due to privacy regulations. The reported work has called in Internet giants such as Baidu, Sougou, Facebook, Yahoo, Huawei and preliminary results have been shown. To be specific, data collection techniques in this work is based on open source solutions of search engines and anonymous Web access logs. Two interesting case studies have been presented. We have a reason to be positive about this benchmark effort considering the big names in the crew.

[Laurila12] also presents a mobile data collection challenge initiated by Nokia, which represents an important step towards mobile big data networking. To be specific, the authors in the work reviewed the Lausanne Data Collection Campaign (LDCC) for unique and longitudinal smartphone data set, which acts as the basis of Mobile Data Challenge (MDC). Privacy, challenging and scalable data collection and usage have been emphasis of this benchmark study.

In sum, remarkable benchmarking efforts have been initiated in both traditional Internet and mobile networking. With an emphasis on privacy-respecting and scalable information collection, the discussed benchmark problems represent a promising step for big data and big networking research in the long run. However, we are also expecting that more insights and inspiring observations can be extracted from these large scale studies.

5.2 Mobile Networking

Mobile networking is becoming a more and more important counterpart of traditional Internet and big data. The mobile networking is becoming larger and larger due to releasing of hundreds of thousands of cell phones and pads. Moreover, the evolution of cellular network has enables mobile devices to be connected fast and reliably.

A number of big data efforts have also been reported regarding mobile networking. [Laurila12] introduced an important mobile big data collection project. [Shekhar12] reported a spatial big-data challenges intersecting mobility and cloud computing. The observation that spatial localization of mobile data is critical is certainly valid and valuable. One interesting improvement for the work in [Shekhar12] is to study daily behaviors of users based on usage of mobile maps on their cellphones or GPS, for which Apple Map and Google Map on cellphones are two important representatives.

A recent effort on mining large-scale smartphone and data for personality studies has been presented in [Chittaranjan13]. Although this work only covers basic aspects of big data, it is still worthwhile a read because it simultaneously considers both personality study and large scale data.

In sum, mobile networking is by fact an important counterpart of traditional Internet. More importantly, benchmarks and case studies have reflected usefulness of studying mobile big data. Moreover, considering the fast and reliable requirement of mobile networking requirements, effective interactions of the cloud and end users (i.e., close-loop control/interaction) might be another interesting research direction.

6. Summary

In this work, we have done in-depth reviews on recent efforts dedicated to big data and big data networking. We have reviewed the progresses in fundamental big data technologies such as storage and warehousing, SDN, transportation and analytics. Important aspects of big data networking in cloud computing such as new challenges and opportunities, resource management and performance optimizations are also introduced and discussed with independent viewpoints. Lastly but not the least, we have also reported important efforts in big data benchmarking and mobile networking, which represent foundations of big data research and promising trends, respectively.

To sum up, we conclude that promising progresses have been made in the area of big data and big data networking, but much remains to be done. Almost all proposed approaches are evaluated at a limited scale, for which the reported benchmarking projects can act as a helpful compensation for larger-scale evaluations. Moreover, software-oriented studies also need to systematically explore cross-layer, cross-platform tradeoffs and optimizations.

Acronyms

SDN Software-Defined Networking

SQL Structured Query Language

NoSQL Not Only SQL

RLE Run Length Encoding

HDFS Hadoop Distributed File System

vSwitch virtual Switch

LDCC Lausanne Data Collection Campaign

MDC Mobile Data Challenge

References

[Laurila12] Laurila, Juha K., et al. The mobile data challenge: Big data for mobile computing research. Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing. 2012. https://research.nokia.com/files/public/MDC2012_Overview_LaurilaGaticaPerezEtAl.pdf

[Costa12] Costa, Paolo, et al. Camdoop: Exploiting in-network aggregation for big data applications. USENIX NSDI. Vol. 12. 2012. http://research.microsoft.com/en-us/um/people/pcosta/papers/costa12camdoop.pdf

[Monga12] Monga, Inder, Eric Pouyoul, and Chin Guok. Software-Defined Networking for Big-Data Science-Architectural Models from Campus to the WAN. High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:. IEEE, 2012. http://www.es.net/assets/pubs_presos/ESnet-SRS-SC12-paper-camera-ready.pdf

[Lakew13] Lakew, Ewnetu Bayuh. Managing Resource Usage and Allocations in Multi-Cluster Clouds. 2013, http://www8.cs.umu.se/~ewnetu/papers/lic.pdf

[Madden12] Madden, Sam. From databases to big data. Internet Computing, IEEE 16.3 (2012): 4-6. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6188576

[Shekhar12] Shekhar, Shashi, et al. Spatial big-data challenges intersecting mobility and cloud computing. Proceedings of the Eleventh ACM International Workshop on Data Engineering for Wireless and Mobile Access. ACM, 2012. http://dmlab.cs.umn.edu/SocialMobileCloud/papers/shashi.pdf

[Chittaranjan13] Chittaranjan, Gokul, Jan Blom, and Daniel Gatica-Perez. "Mining large-scale smartphone data for personality studies." Personal and Ubiquitous Computing 17.3 (2013): 433-450. http://publications.idiap.ch/downloads/papers/2011/Chittaranjan_PUC_2012.pdf

[Girola11] Girola, Michele, et al. "IBM Data Center Networking: Planning for virtualization and cloud computing." GOOGLE/IP. COM/IBM Redbooks (2011). http://www.redbooks.ibm.com/redbooks/pdfs/sg247928.pdf

[Brunet12] Brunet, Pierre-Marie, Alain Montmorry, and Benoît Frezouls. "Big data challenges, an insight into the Gaia Hadoop solution." (2012). http://www.spaceops2012.org/proceedings/documents/id1275512-Paper-003.pdf

[Keim13] Keim, Daniel, Huamin Qu, and Kwan-Liu Ma. "Big-Data Visualization." Computer Graphics and Applications, IEEE 33.4 (2013): 20-21. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6562707

[Dai11] Dai, Jinquan, et al. "Hitune: dataflow-based performance analysis for big data cloud." Proc. of the 2011 USENIX ATC (2011): 87-100. https://www.usenix.org/legacy/event/atc11/tech/final_files/Dai.pdf

[Gao13] Gao, Wanling, et al. "Bigdatabench: a big data benchmark suite from web search engines." arXiv preprint arXiv:1307.0320 (2013), http://arxiv.org/pdf/1307.0320.pdf

[Ferguson12] Ferguson, Mike. "Architecting A Big Data Platform for Analytics." A Whitepaper Prepared for IBM (2012). http://public.dhe.ibm.com/common/ssi/ecm/en/iml14333usen/IML14333USEN.PDF

[Ji12] Ji, Changqing, et al. "Big data processing in cloud computing environments." Pervasive Systems, Algorithms and Networks (ISPAN), 2012 12th International Symposium on. IEEE, 2012. http://192.240.0.102/downloads/MAG/vol48-2/paper09.pdf

[Lu11] Lu, Sifei, et al. "A framework for cloud-based large-scale data analytics and visualization: Case study on multiscale climate data." Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference on. IEEE, 2011. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6133204&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6133204

[Budiu12] Budiu, Mihai. "Putting a big-data platform to good use: training kinect." Proceedings of the 21st international symposium on High-Performance Parallel and Distributed Computing. ACM, 2012. http://budiu.info/work/hpdc12.pdf

[Zhang13] Zhang, Linquan, et al. "Moving Big Data to The Cloud: An Online Cost-Minimizing Approach." IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS 31.12 (2013): 1. http://i.cs.hku.hk/~fcmlau/papers/info13-lq-m.pdf

[Wang12] Wang, Guohui, T. S. Ng, and Anees Shaikh. "Programming your network at run-time for big data applications." Proceedings of the first workshop on Hot topics in software defined networks. ACM, 2012. http://www.cs.rice.edu/~eugeneng/papers/HotSDN12.pdf

[Sadasivarao13] Sadasivarao, Abhinava, et al. "Bursting Data between Data Centers: Case for Transport SDN." High-Performance Interconnects (HOTI), 2013 IEEE 21st Annual Symposium on. IEEE, 2013. http://ieeexplore.ieee.org/iel7/6626675/6627714/06627742.pdf?arnumber=6627742

[Agrawal11] Agrawal, Divyakant, Sudipto Das, and Amr El Abbadi. "Big data and cloud computing: current state and future opportunities." Proceedings of the 14th International Conference on Extending Database Technology. ACM, 2011. http://www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf

[Herodotou11] Herodotou, Herodotos, et al. "Starfish: A Self-tuning System for Big Data Analytics." CIDR. Vol. 11. 2011. http://www.cs.duke.edu/~gang/documents/CIDR11_Paper36.pdf

[He11] He, Yongqiang, et al. "RCFile: A fast and space-efficient data placement structure in MapReduce-based warehouse systems." Data Engineering (ICDE), 2011 IEEE 27th International Conference on. IEEE, 2011. http://www.cse.ohio-state.edu/hpcs/WWW/HTML/publications/papers/TR-11-4.pdf

[Fisher12] Fisher, Danyel, et al. "Interactions with big data analytics." interactions 19.3 (2012): 50-59. http://research.microsoft.com/pubs/163593/inteactions_big_data.pdf

[Begoli12] Begoli, Edmon, and James Horey. "Design Principles for Effective Knowledge Discovery from Big Data." Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on. IEEE, 2012.http://cda.ornl.gov/publications_2012/Publication_36116.pdf

[Prekopcsák11] Prekopcsák, Zoltán, et al. "Radoop: Analyzing big data with rapidminer and hadoop." Proceedings of the 2nd RapidMiner Community Meeting and Conference (RCOMM 2011). 2011. http://prekopcsak.hu/papers/preko-2011-rcomm.pdf

[Dittrich12] Dittrich, Jens, and Jorge-Arnulfo Quiané-Ruiz. "Efficient big data processing in Hadoop MapReduce." Proceedings of the VLDB Endowment 5.12 (2012): 2014-2015. http://vldb.org/pvldb/vol5/p2014_jensdittrich_vldb2012.pdf

[Silberstein11] Silberstein, Adam, Ashwin Machanavajjhala, and Raghu Ramakrishnan. "Feed following: the big data challenge in social applications." Databases and Social Networks. ACM, 2011.

http://www.cs.duke.edu/~ashwin/pubs/FeedFollowing_DBSocial11.pdf

[Bakshi12] Bakshi, Kapil. "Considerations for big data: Architecture and approach." Aerospace Conference, 2012 IEEE. IEEE, 2012. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6187357

[Zaslavsky13] Zaslavsky, Arkady, Charith Perera, and Dimitrios Georgakopoulos. "Sensing as a service and big data." arXiv preprint arXiv:1301.0159 (2013), http://arxiv.org/pdf/1301.0159v1.pdf

[Kaushik12] Kaushik, Rini T., and Klara Nahrstedt. "T: a data-centric cooling energy costs reduction approach for big data analytics cloud." Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. IEEE Computer Society Press, 2012. http://conferences.computer.org/sc/2012/papers/1000a037.pdf

[Cisco11]Cisco White Paper, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2010-2015, Feb. 2011.http://newsroom.cisco.com/ekits/Cisco_VNI_Global_Mobile_Data_Traffic_Forecast_2010_2015.pdf

---

Last Modified: December 10, 2013
This and other papers on latest advances in computer networking are available on line at
http://www.cse.wustl.edu/~jain/cse570-13/index.html
Back to Raj Jain's Home Page