ENERGY CONSUMPTION IN BIG DATA ENVIRONMENTS– A SYSTEMATIC MAPPING STUDY

Erlend Pettersen

Østfold University College

Halden 1783, Norway

erlendp@hiof.no

Ricardo Colomo-Palacios

Østfold University College

Halden 1783, Norway

ricardo.colomo-palacios@hiof.no

Big Data is a term that describes a large volume of structured and unstructured data. Big Data must be acquired, stored, analyzed and visualized by means of non-conventional methods requiring normally a big set of resources, which includes energy consumption. Although Big Data is not new as a phenomenom, its explosion of the interest in literature is recent and its study in new scenarios presents several gaps. On the other hand, Green IT is also a growing field in computing, given the increasing role of IT in energy consumption in the world. Green IT is aimed to reduce IT-related energy consumption and overall IT environmental impact. In order to investigate the reported initiatives regarding the Big Data and Green IT with a focus of energy consumption, the authors conducted a systematic mapping on the topic. The search strategy which was used resulted in 28 relevant studies which were relevant to the topic. We found that a majority of the studies performed present algorithms designed to reduce the energy consumption in data centres. The rest of the studies present benchmarks and energy measurements, reviews, proposals of hardware-based solutions, as well as studies which give an overview of one or more aspects on Big Data.

1 INTRODUCTION

The use of cloud computing and Big Data enables the analysis and processing of massive amounts of structured and unstructured data. The analysis of this data presents many applications, such as in entertainment, medicine or science, citing just some of the most important features (E. Casalicchio, Lundberg, & Shirinbab, 2016; Emiliano Casalicchio, Lundberg, & Shirinbab, 2017; Niemi, Nurminen, Liukkonen, & Hameri, 2018; Vera-Baquero, Colomo-Palacios, & Molloy, 2016) . The amount of collected and generated data is increasing. The terms Volume, Variety and Velocity are used to refer to the amount, the variety of data (from various sources), and to the speed of which data is generated (Zikopoulos, 2012). The data centers which are necessary to implement Big Data technologies require large amounts of energy, and is estimated to use about 8% of the worldwide electricity by 2020 (Yang, Pen, Li, & Zomaya, 2017). This causes an increase in carbon emissions. The authors propose a systematic mapping on energy consumption in Big Data environments. Our main motivation for performing this study is double. On the one hand, the combination of big data with energy consumption is a hot topic nowadays and on the other hand, to the best of our knowledge, there is not a systematic mapping on this topic published in the literature.

Cloud computing is the utilization of a shared pool of configurable computing resources. The cloud service provider’s computing resources are pooled to serve multiple customers. The available resources may be processing, storage, network and memory bandwidth. The resources may be provisioned or released on demand with minimal management effort or interaction with the service provider (Peter McNell & Tim Grance, 2018).

A Service Level Agreement (SLA) is an agreement between the cloud service provider and the customer. The agreement commonly contains a definition of the services provided, a specification of the performance level of the services offered, how to resolve unplanned incidents, warranties and remedies, and a definition of the duties and reponsibilites of the customers (Kandukuri, V., & Rakshit, 2009). QoS (Quality of Service) is a measure of properties such as response time, throughput, and failure probability, as experienced by the users of a cloud service (Zheng, Wu, Zhang, Lyu, & Wang, 2013). A Virtual machine is a concept where an an isolated environment is created on a physical machine. The physical machine may host more than one virtual machine simultaneously. Applications run inside each virtual machine (Rosenblum, 2004). The use of virtual machines may cause a decrease in the total energy usage of a data centre, as several virtual machines may be consolidated on a small set of physical machines (Altomare, Cesario, & Vinci, 2018). An algorithm consists of a set of steps which defines how a task is to be performed. Each step of an algorithm must be a unambigious (Brookshear, Smith, & Brylow, 2012). In Big Data processing, a scheduling algorithm is an algorithm which attempts to achieve the proper utilization of resources (Gautam, Prajapati, Dabhi, & Chaudhary, 2015). MapReduce is both a programming model and an implementation for processing and generating large data sets. The users specifies the computation in terms of a map- and a reduce-function. The underlying system automatically parallelizes the computation across a cluster of machines. This has applications in fields such as data mining, machine learning, text processing (Dean & Ghemawat, 2008). Hadoop is an open-source implementation of Hadoop which was originally developed at Yahoo (Zikopoulos, 2012). The data and the computation work is distributed among a set of servers (also known as a cluster). Hadoop utilizes the Hadoop Distributed File System (HDFS) for accessing and writing data, and for performing the replication of data among the nodes in a cluster (Shvachko, Kuang, Radia, & Chansler, 2010).

The rest of this paper is organized as follows: In section two, the research methodology, research questions, study selection, classification, and the data extraction process is presented. Subsequently, we present our analysis of the results of the systematic mapping. Finally, main conclusions are depicted and future work presented.

2 RESEARCH METHODOLOGY

In this section we present the research methodology, research questions, search strategy, the study selection (including the inclusion and exclusion criteria) and study classification, and we describe the data extraction process.

The focus of this study is to give an overview of the current state of research regarding energy consumption and Big Data combined as a field of study. To be able to do this, a Systematic Mapping was conducted. The study was performed in accordance to a set of guidelines for performing Systematic Mappings (Petersen, Feldt, Mujtaba, & Mattsson, 2008).

The final output of the research process is a systematic mapping of the current research conducted on the topic of Big Data and Green IT with a focus on energy consumption.

2.1 Research Questions

In order to be able to perform the study, it was necessary to define a set of research questions. The following research questions were formulated:
RQ1: What types of research studies have been published on the topic of Big Data and Green IT?
RQ2: What have the published studies reported regarding the main effects and influencers on energy consumption within Big Data?
RQ3: In which publishing outlets have relevant studies been published?

2.2 Search strategy

The following databases were used to find relevant literature:

- ACM Digital Library (http://dl.acm.org)
- IEEE Explore (https://explore.ieee.com)
- SpringerLink (https://link.springer.com)
- Taylor & Francis (https://www.tandfonline.com/)
- Wiley Online Library (https://onlinelibrary.wiley.com)

These databases were chosen as they are among the most relevant sources of articles within the broad field of computing. The term “big data”, “energy consumption”, “power consumption”, “green IT”, and “green computing” were considered to be the primary terms. Relevant synonyms and terms which have similar meanings as the primary terms were considered. The terms and synonyms which were considered relevant were added to the list of terms. This was done in order to ensure it would be possible to find a sufficient amount of studies.

It was decided to use the following terms in the search string: Big Data, energy consumption, power consumption, Green IT, Green Information Systems, Green IS, IT for Green, IT Energy Management, Green ICT, Green Computing. In accordance with the research questions, authors wanted to find studies which were on the topic of Big Data and energy consumption. This is reflected in the search string and in the inclusion criteria. The following search string was used:

- ("Big Data" AND ("energy consumption" OR "power consumption") AND ("Green IT" OR "Green Information Systems" OR "Green IS" OR "IT for Green" OR "IT Energy Management" OR "Green ICT" OR "Green Computing"))

The search string was entered into the search field of the library databases mentioned above. When the search was performed on the ACM Digital Library, the search string had to be rewritten to fit the requirements of the search interface of this database. For all of the results, the titles and keywords were reviewed for relevancy. Articles which were deemed irrelevant were discarded. The remaining articles were saved in a reference management system for further review. The titles, abstracts and keywords of the saved articles were read, and checked against the inclusion and exclusion criteria. The full text of the articles which passed the inclusion criteria were read. In some cases, the title, abstract and the keywords of an article passed the inclusion criteria, but upon reading the full text of an article, it was apparent that the article did not pass the criteria. When this was the case, the article was rejected.

2.3 Study Selection

A study selection is done in order to be able to find papers which have relevance to the research questions. A key criteria was that a study must have had a focus on Big Data. The following inclusion criteria were formulated:

- The title, the abstract, or the list of keywords for an article must contain the term “Big Data”.
- In addition, the title, the abstract, or the list of keywords must contain at least one of the following words: “energy consumption, “power consumption”, “Green IT”, “Green Information Systems”, “Green IS”, “IT for Green”, “IT Energy Management”, “Green ICT”, “Green Computing”.
- The paper must have relevance to the topic of Big Data, and energy or environmental issues.
- The paper is an academic work, which has been published in a journal or in a publication of proceedings.
- The paper is available.

Exclusion criteria were as follows:

- *Abstract:* The abstract states that the article is on the subject of smart grids and electricity grids and not on Big Data as the field of study.
- *Accessibility:* The study is not available through the libraries which Østfold University College has access to.
- *Language:* The study has not been published in the English language.
- *Title and abstract:* The title and/or the abstract does not match the inclusion criteria.
- *Full text:* After having read the full text of an article, it is apparent that the content has no relevancy to the research questions.

The studies were collected by April, 2018.

2.4 Study Classification

The collected papers were categorized, based on the title, abstract, and the keywords encountered in the papers, and the full text. Based on this work, the following categories were determined:

- Algorithms / models: Studies which proposes mathematical algorithms and/or models which can be used to calculate and/or optimize the energy consumption or energy efficiency of servers. Studies of this type may also contain simulations developed by the researchers, and benchmarks of the algorithm. However, the main focus in this type of study is on the proposed algorithm or model.
- Measurements of energy efficiency (benchmarking): Studies where the primary focus is on measurement and comparisation of the energy consumption of servers, networking equipment, or other types of hardware. The measurements are generally done using various scenarios and usage patterns.
- Hardware: Articles which propose the use of hardware devices to help reduce the energy consumption.
- Review: Articles which presents a review of current practices within the field, or gives an overview of one or more aspects related to Big Data, energy consumption and/or related environmental issues.

2.5 Data Extraction

In order to evaluate whether each collected paper would have relevancy to answer the research questions, a table was created. This table presents columns containing details about the title of the papers, where (in which database) the paper was encountered, and details about whether the paper passed the inclusion criteria. Additionally, a column was created with notes describing why a paper was included or rejected. Authors found that three authors have published two articles (Maroulis, Zacheilas, & Kalogeraki, 2017a, 2017b). The year with the largest amount of published articles is 2016, when 11 articles were published. The collected articles were published in the years 2011-2018, with 2016 (11) and 2017 (7) being the years when the majority of the articles were published.

3 ANALYSIS AND DISCUSSION OF RESULTS

In this section, the results and the findings of the systematic mapping study will be presented. All of the encountered papers were checked against the inclusion and exclusion criteria. 38 papers were selected, based on the title, the abstract and the keywords. The full text of these articles were read. 28 of the papers were found relevant to the research questions.

The numbers reported in parentheses in Table 1 are the number of articles which were returned by the given database after specifying that only freely available studies should be listed by the library database. This was necessary with Springer Link and Taylor & Francis.

Library	**Total number of results**	**Number of papers selected based on title, abstract and/or keywords**	**Number of papers selected based on the full text (content)**
ACM	73	0	0
IEEE Explore	35	20	14
Science Direct	980	12	9
Springer Link	144 (62 accessible)	5	4
Taylor & Francis	12 (11 accessible)	0	0
Wiley Online Library	25	1	1
Results	**1269 (1186)**	**38**	**28**

Table 1: Table of libraries, and the results of the filtering process

3.1 RQ1: What types of research studies have been published on the topic of Big Data and Green IT?

Paper	**Algorithms / models**	**Benchmarking / measurements**	**Hardware**	**Review**
(Emiliano Casalicchio et al., 2017)	X			
(Godbole & Lamb, 2015)				X
(Gürbüz & Tekinerdogan, 2016)				X
(Ismail & Fardoun, 2016)	X			
(Marotta, Avallone, & Kassler, 2018)	X			
(Niemi et al., 2018)				X
(Shu & Wu, 2017)	X			
(Shuja et al., 2017)				X
(Song, He, Wang, Yu, & Pierson, 2016)	X			
(Tran, Do, Rotter, & Hwang, 2018)	X			
(J. Wu, Guo, Li, & Zeng, 2016)				X
(Zhai et al., 2016)	X			
(Al-Salim, Ali, Lawey, El-Gorashi, & Elmirghani, 2016)	X			
(Camp & Chauveau, 2017)		X		
(Ho & Pernici, 2015)	X			
(Kaushik, Abdelzaher, Egashira, & Nahrstedt, 2011)	X			
(Maroulis et al., 2017a)	X			
(Maroulis et al., 2017b)	X			
(Nabavinejad & Goudarzi, 2016)	X			
(Wei & Ren, 2014)				X
(Xue et al., 2017)	X			
(Dou Wanchun et al., 2016)	X			
(Rahman & Esmailpour, 2016)			X	
(Rong, Zhang, Xiao, Li, & Hu, 2016)				X
(Shao, Li, Gu, Zhang, & Luo, 2018)	X			
(W. Wu, Lin, Hsu, & He, 2017)				X
(Yang et al., 2017)	X			
(Zong, Ge, & Gu, 2017)			X	
SUM	17	1	2	8

Table 2: The results of the categorization of the selected studies

The studies were categorized into four distinct categories, based on the title, keywords, abstract, and the full text. The categories are described in section 2.4. Table 2 lists the results of the categorization of the studies.

The results shows that a majority of the published studies were proposals of algorithms and models which aims to reduce the energy consumption while Big Data analysis is taking place. The second largest category is reviews, which contains studies with tertiary studies of one or more aspects related to Big Data and energy consumption.

3.2 RQ2: What have the published studies reported regarding the main effects and influencers on energy consumption within Big Data?

Topic:	**Study:**
Energy efficiency in scientific computing	(Niemi et al., 2018)
Energy aware auto scaling	(Emiliano Casalicchio et al., 2017)
Formulation of a workflow mapping problem, a proposal for a fully polynomial-time approximation scheme, and the results of a simulation	(Shu & Wu, 2017)
Literature review of green computing metrics, which focuses on parallel computing and green computing of Big Data systems	(Gürbüz & Tekinerdogan, 2016)
Literature review on ways to optimize the energy consumption on Hadoop clusters	(W. Wu et al., 2017)
Optimal distribution of tasks, in order to reduce the energy consumption	(Ismail & Fardoun, 2016; Marotta et al., 2018; Song et al., 2016; Tran et al., 2018)
Use of an energy efficient virtual machine placement and route scheduling scheme to reduce the energy consumption	(Yang et al., 2017)
Overview of research on Big Data and energy consumption	(J. Wu et al., 2016)
Review of Big Data and Green IT, and its applications in healthcare	(Godbole & Lamb, 2015)
Scheduling algorithms	(Dou Wanchun et al., 2016; Maroulis et al., 2017a, 2017b; Shao et al., 2018; Xue et al., 2017; Zhai et al., 2016)
Review of the Green Computing paradigm, with a focus on emerging IT technologies	(Shuja et al., 2017)
Use Processing Nodes to preprocess data, to reduce the amount of data to be transferred	(Al-Salim et al., 2016)
Comparisation of the energy cost of two ways to insert data into a Oracle database	(Camp & Chauveau, 2017)
Adaptation framework, to obtain energy efficiency	(Ho & Pernici, 2015)
Use machine learning to create zone placement, migration and replication policies	(Kaushik et al., 2011)
Scheme to find the best virtual machine configuration	(Nabavinejad & Goudarzi, 2016)
Overview of the energy consumption of current data centres	(Rong et al., 2016; Wei & Ren, 2014)
Networking infrastracture in a data centre	(Rahman & Esmailpour, 2016)
A system built to perform research in energy-aware high performance computing and big data analytics	(Zong et al., 2017)

Table 3: Table of topics encountered in the selected papers.

Several studies have considered the topic of scheduling of workloads. An overloaded server causes a reduction in the energy efficiency. When too many tasks are running simultaneously, they take more time to complete. Servers also consume electricity when they are not performing any useful work, i.e. when they are idle. One study proposes an algorithm which aims to reduce energy consumption and to increase energy efficiency by performing an optimal distribution of divisible tasks between servers, in order to ensure a high utilization of the available servers. The individual tasks must not be dependent on synchronization with other tasks, and they must not perform any inter-task communication (Ismail & Fardoun, 2016). Somewhat similarly, a study proposed an algorithm for consolidating workloads on as few physical servers as possible (Marotta et al., 2018). One study was performed on the scheduling of workloads on Apache Spark clusters, in order to minimize the energy consumption. The proposed scheduler performs energy-efficient scheduling, using an EDF scheduling policy and by varying the CPU frequency while the applications are executing on the cluster (Maroulis et al., 2017a). The same authors proposed another scheduling algorithm, also on Apache Spark, which attempts to find the appropriate number of worker nodes for an application, and which uses DVFS (Dynamic Voltage and Frequency Scaling) to set the CPU frequency. The proposed scheduling algorithm determines the amount of resources to to allocate per submitted application, in order to satisfy both performance requirements and to minimize the energy consumption (Maroulis et al., 2017b). (Shu & Wu, 2017) investigated the property of moldable jobs and formulated a workflow mapping problem to minimize the dynamic energy consumption under deadline and resource constraints. The results of the simulation and the results of the real-life workflow implementation on Hadoop and YARN-systems shows that the FPTAS (fully polynomial-time approximation scheme) saves energy. (Tran et al., 2018) proposes an algorithm for scheduling which takes into account the performance of each individual server. The algorithm is more energy efficient than the default layout scheme used by Hadoop. (Zhai et al., 2016) proposed the “Green Scheduler” for use with Hadoop. Benchmarks performed by the researchers shows that using the Green Scheduler results in less consumption of energy compared to when the FIFO scheduler or the Capacity scheduler is used.

Finding the best VM configuration is an important issue when efficient execution of MapReduce-jobs is desired. The Smart Configuration Selection (SCS) scheme attempts to find the most suitable sample configurations, by running the application on a few configurations, and then to estimate the missing values. The results shows that the scheme finds the best configuration on average, but it cannot supply the best answer for each specific application (Nabavinejad & Goudarzi, 2016).

One study reports that it is possible to characterize Big Data workloads, and to find the data value based on the levels of usefulness and importantness of the data. This information can be used with an adaptation framework to obtain energy efficiency. The approach attempts to maximime the throughput of high-value workloads (Ho & Pernici, 2015).

A study found that in a Hadoop cluster, the file size, file lifespan and file heat are statistically correlated and strongly associated with the directory structure. The study proposes that by using supervised machine learning, it is possible to implement predictive file zone replacement, migration and replication policies. The researchers calls their approach Predictive GreenHDFS. The approach was tested with a simulation, using real data from a large-scale production Hadoop cluster. The results shows that Predictive GreenHDFS results in higher energy savings, performance and greater storage savings compared to when Reactive GreenHDFS is used (Kaushik et al., 2011).

Two studies investigated the issue of energy consumption QoS, and proposed QoS-aware energy scheduling algorithms. An algorithm was proposed to perform VM-migration between servers, with a focus on QoS. This algorithm aims to enhance price and execution time, during the execution of data-intensive computation tasks. A simulation was performed, which shows that the proposed algorithm is works as intended. (Dou Wanchun et al., 2016). Another QoS-based energy aware scheduling algorithm attempts to achieve energy savings during task scheduling, based on QoS. The algorithm takes into account the power consumption of the servers (physical machines) and the task requirements (Xue et al., 2017).

Two studies have considered the issues of energy consumption and SLA (Service Level Agreements). In a study on Apache Cassandra virtual data centres, an optimal adapation model was proposed to handle these issues. The adaptation strategies used were horizontal scaling, vertical scaling, and optimal placement (Emiliano Casalicchio et al., 2017). Another study considered energy consumption and meeting the users’ SLA when using Hadoop-based systems. The authors proposed a framework which consists of a enery-aware fair based scheduling, and a manager for turning nodes off and on (Shao et al., 2018).

A study proposed an algorithm to improve the parallelism of a MapReduce system. This is done by ensuring that the mapping and reduction process is completed simultaneously, by distributing the data blocks to the nodes based on their capabilities and the data features (Song et al., 2016).

A study on mass-insertion of data into Oracle-based databases, showed that using the FORALL-method gives better results when response time and energy consumption is considered (Camp & Chauveau, 2017).

The Marcher-system may be used by researchers to perform research on energy consumption, or for education purposes. The system consists of sensors which measures the power of the different components of a computer. The system is available to researchers via the Web or by using SSH. Marcher supports a range of widely used programming languages and most parallel programming models. The system gives detailed performance and power profiles for computer systems and components (Zong et al., 2017).

One study was performed on the placing of virtual machines (VMs), using VM migration technology and data flow consolidation technology to minimize the energy consumption and to balance the communication load within the data center network. Simulations performed by the researchers shows that the proposed JAVRS (Job-Aware VM Placement and Route Scheduling) algorithm outperforms two well-known QAP algorithms, uses less energy consumption and enables better network performance (Yang et al., 2017).

(Rahman & Esmailpour, 2016) proposes the use of optical networking equipment for parts of the infrastructure in a data centre. Big Data causes a large volume of operations, and traditional Ethernet-based links are not sufficient to handle the resulting traffic. The study proposes using both Ethernet and optical based links to connect the servers with the upper layer in the data centre. Simulations shows that with the proposed solution, there is less latency and higher throughput compared to when Ethernet-based links are used. Optical links uses less electricity, and the need for cooling is reduced.

Processing Nodes (PN), a node which is capable of processing and storing data, may be used to process data as much as possible, before the result is transferred to a data centre for further processing. This may be done to reduce the amount of data to be transferred to the data centre, and reduces the power consumption. In addition, the authors describes the Mixed Integer Linear Programming Model (MILP) (Al-Salim et al., 2016).

Several of the studies contained reviews of one or more aspects related to Big Data and energy consumption. One study gives a brief overview of current benchmarks for energy efficiency. The study covers power and energy measurements, benchmarking and analysis in big data processing. (Wei & Ren, 2014). An extensive review was performed on studies performed on Hadoop and ways to optimize the energy consumption. The study cites a large amount of studies performed on this topic, and categorizes the reviewed studies into five categories (W. Wu et al., 2017). Another extensive review discusses the subject of optimization of energy consumption for data centers. The study reports that servers and the refrigeration (cooling) systems consumes the most energy in a data centre. A general energy saving framework is presented, and the authors reviewed the main progress of the proposed energy conservation technology, introduced energy consumption evaluation methods, and conducted a side-by-side comparisation of their accuracy and efficiency (Rong et al., 2016). (Gürbüz & Tekinerdogan, 2016) gives a review of software metrics and Green IT, with a focus on parallel computing, and uses existing literature reviews and systematic mapping studies to identify challenges and and operating big data systems. The researchers found that existing studies mainly had considered metrics such as speedup and efficiency. The study also discusses the green computing metrics that are of interest of the analysis of energy efficient deployment alternatives. (Shuja et al., 2017) conducted a review of several topics related to green computing. Big Data was one of the subjects. In (J. Wu et al., 2016) a comprehensive literature survey was performed on how to green big data systems in terms of the whole life cycle of big data processing. The study proposes two new metrics.

(Godbole & Lamb, 2015) conducted a review of the use of data science and big data to reduce the environmental impact of healthcare. Some of the suggestions includes deduplication of data, use virtualization, and to use electronic medical records (EMR).

Finally, a study performed on the topics of scientific computing, Big Data, and energy consumption at CERN reported that the issue of energy consumption became a factor when the amount of available power limited the number of new servers which could be installed in the data centre. The study also reports that end users, software developers and the data centre managers have different views on how to achieve higher energy efficiency. In scientific computing there is a high threshold to perform modifications to thoroughly tested and validated code to gain better performance. Due to differences in the way floating point calculcations are implemented in compilers and operating systems, the scientific results produced may not be correct after the modifications have been implemented in the code (Niemi et al., 2018).

Channel	#	Title
Journal (15)	4	Future Generation Computer Systems
	2	Big Data Research
	2	Journal of Grid Computing
	1	Cluster Computing
	1	Computers & Industrial Engineering
	1	Concurrency and Computation: Practice and Experience
	1	IEEE Systems Journal
	1	Journal of Internet Services and Applications
	1	Procedia Computer Science
	1	Renewable and Sustainable Energy Reviews
Conference	1	2011 International Green Computing Conference and Workshops
	1	2014 23rd International Conference on Computer Communication and Networks (ICCCN)
	1	2015 12th International Conference Expo on Emerging Technologies for a Smarter World (CEWIT)
	1	2015 IEEE 8th International Conference on Cloud Computing
	1	2015 IEEE Conference on Technologies for Sustainability (SusTech)
	1	2016 18th International Conference on Transparent Optical Networks (ICTON)
	1	2016 Design, Automation Test in Europe Conference Exhibition (DATE)
	1	2016 IEEE First International Conference on Data Science in Cyberspace (DSC)
	1	2016 IEEE International Congress on Big Data (BigData Congress)
	1	2016 International Conference on Cloud and Autonomic Computing (ICCAC)
	1	2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)
	1	2017 IEEE International Conference on Autonomic Computing (ICAC)
	1	2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

Table 4: Publishing outlets

3.3 RQ3: In which publishing outlets have relevant studies been published?

The journal “Future Generation Computer System” published four articles on this topic. The journal has, among other topics, a focus on Big Data registration, processing and analyses. The other journals with more than one relevant article were “Big Data Research” (2) and “Journal of Grid Computing (2). The rest of the publishing outlets are listed in Table 4.

4 CONCLUSION AND FUTURE WORK

In this systematic mapping, the current state of Big Data and Green IT with focus on energy consumption is outlined. This was achieved by identifying the relevant studies, and performing an analysis and categorization of these studies. The majority of the studies proposed algorithms which causes a reduction in the energy consumption when Big Data is being performed in data centres.

Among the algorithms, authors find that a majority of the studies proposes various scheduling algorithms. The second largest category contained reviews of various aspects related to Big Data.

One selected study reported that there is a symbiotic relationship between cloud computing and Big Data (Godbole & Lamb, 2015). A future work may be to perform a wider systematic mapping study or a literature review on these topics but also conduct a research study on the intersection of these knowledge areas. Another future work may be to perform a systematic mapping on Big Data and energy consumption, which also includes the commonly used frameworks and platforms in the search strings. This may give more relevant results, and may give more insight on this topic. Finally, it is aimed to investigate the use of GPUs to reduce the computation time in Big Data applications.

5 REFERENCES

Al-Salim, A. M., Ali, H. M. M., Lawey, A. Q., El-Gorashi, T., & Elmirghani, J. M. H. (2016). Greening big data networks: Volume impact. In *2016 18th International Conference on Transparent Optical Networks (ICTON)* (pp. 1–6). https://doi.org/10.1109/ICTON.2016.7550707

Altomare, A., Cesario, E., & Vinci, A. (2018). Data analytics for energy-efficient clouds: design, implementation and evaluation. *International Journal of Parallel, Emergent and Distributed Systems*, 1–16. https://doi.org/10.1080/17445760.2018.1448931

Brookshear, J. G., Smith, D. T., & Brylow, D. (2012). *Computer science: an overview* (11th ed). Boston: Addison-Wesley.

Camp, O., & Chauveau, E. (2017). Measuring the Energy Consumption of Massive Data Insertions: An Energy Consumption Assessment of the PL/SQL FOR LOOP and FORALL Methods. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (pp. 450–457). https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.73

Casalicchio, E., Lundberg, L., & Shirinbab, S. (2016). Energy-Aware Adaptation in Managed Cassandra Datacenters. In *2016 International Conference on Cloud and Autonomic Computing (ICCAC)* (pp. 60–71). https://doi.org/10.1109/ICCAC.2016.12

Casalicchio, E., Lundberg, L., & Shirinbab, S. (2017). Energy-aware auto-scaling algorithms for Cassandra virtual data centers. *Cluster Computing*, *20*(3), 2065–2082. https://doi.org/10.1007/s10586-017-0912-6

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107. https://doi.org/10.1145/1327452.1327492

Dou Wanchun, Xu Xiaolong, Meng Shunmei, Zhang Xuyun, Hu Chunhua, Yu Shui, & Yang Jian. (2016). An energy-aware virtual machine scheduling method for service QoS enhancement in clouds over big data. *Concurrency and Computation: Practice and Experience*, *29*(14), e3909. https://doi.org/10.1002/cpe.3909

Gautam, J. V., Prajapati, H. B., Dabhi, V. K., & Chaudhary, S. (2015). A survey on job scheduling algorithms in Big data processing (pp. 1–11). IEEE. https://doi.org/10.1109/ICECCT.2015.7226035

Godbole, N. S., & Lamb, J. (2015). Using data science big data analytics to make healthcare green. In *2015 12th International Conference Expo on Emerging Technologies for a Smarter World (CEWIT)* (pp. 1–6). https://doi.org/10.1109/CEWIT.2015.7338161

Gürbüz, H. G., & Tekinerdogan, B. (2016). Software Metrics for Green Parallel Computing of Big Data Systems. In *2016 IEEE International Congress on Big Data (BigData Congress)* (pp. 345–348). https://doi.org/10.1109/BigDataCongress.2016.54

Ho, T. T. N., & Pernici, B. (2015). A data-value-driven adaptation framework for energy efficiency for data intensive applications in clouds. In *2015 IEEE Conference on Technologies for Sustainability (SusTech)* (pp. 47–52). https://doi.org/10.1109/SusTech.2015.7314320

Ismail, L., & Fardoun, A. (2016). EATS: Energy-Aware Tasks Scheduling in Cloud Computing Systems. *Procedia Computer Science*, *83*, 870–877. https://doi.org/10.1016/j.procs.2016.04.178

Kandukuri, B. R., V., R. P., & Rakshit, A. (2009). Cloud Security Issues (pp. 517–520). IEEE. https://doi.org/10.1109/SCC.2009.84

Kaushik, R. T., Abdelzaher, T., Egashira, R., & Nahrstedt, K. (2011). Predictive data and energy management in GreenHDFS. In *2011 International Green Computing Conference and Workshops* (pp. 1–9). https://doi.org/10.1109/IGCC.2011.6008563

Marotta, A., Avallone, S., & Kassler, A. (2018). A Joint Power Efficient Server and Network Consolidation approach for virtualized data centers. *Computer Networks*, *130*, 65–80. https://doi.org/10.1016/j.comnet.2017.11.003

Maroulis, S., Zacheilas, N., & Kalogeraki, V. (2017a). A Framework for Efficient Energy Scheduling of Spark Workloads. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)* (pp. 2614–2615). https://doi.org/10.1109/ICDCS.2017.179

Maroulis, S., Zacheilas, N., & Kalogeraki, V. (2017b). ExpREsS: EneRgy Efficient Scheduling of Mixed Stream and Batch Processing Workloads. In *2017 IEEE International Conference on Autonomic Computing (ICAC)* (pp. 27–32). https://doi.org/10.1109/ICAC.2017.43

Nabavinejad, S. M., & Goudarzi, M. (2016). Energy efficiency in cloud-based MapReduce applications through better performance estimation. In *2016 Design, Automation Test in Europe Conference Exhibition (DATE)* (pp. 1339–1344).

Niemi, T., Nurminen, J. K., Liukkonen, J.-M., & Hameri, A.-P. (2018). Towards Green Big Data at CERN. *Future Generation Computer Systems*, *81*, 103–113. https://doi.org/10.1016/j.future.2017.11.001

Peter McNell, & Tim Grance. (2018, September). The NIST Definition of Cloud Computing. National Institute of Standards and Technology. Retrieved from https://csrc.nist.gov/publications/detail/sp/800-145/final

Petersen, K., Feldt, R., Mujtaba, S., & Mattsson, M. (2008). Systematic Mapping Studies in Software Engineering, (12th International Conference on Evaluation and Assessment in Software Engineering (EASE)), 68–77.

Rahman, M. N., & Esmailpour, A. (2016). A Hybrid Data Center Architecture for Big Data. *Big Data Research*, *3*, 29–40. https://doi.org/10.1016/j.bdr.2016.02.001

Rong, H., Zhang, H., Xiao, S., Li, C., & Hu, C. (2016). Optimizing energy consumption for data centers. *Renewable and Sustainable Energy Reviews*, *58*, 674–691. https://doi.org/10.1016/j.rser.2015.12.283

Rosenblum, M. (2004). The Reincarnation of Virtual Machines. *Queue*, *2*(5), 34–40. https://doi.org/10.1145/1016998.1017000

Shao, Y., Li, C., Gu, J., Zhang, J., & Luo, Y. (2018). Efficient jobs scheduling approach for big data applications. *Computers & Industrial Engineering*, *117*, 249–261. https://doi.org/10.1016/j.cie.2018.02.006

Shuja, J., Ahmad, R. W., Gani, A., Ahmed, A. I. A., Siddiqa, A., Nisar, K., … Zomaya, A. Y. (2017). Greening emerging IT technologies: techniques and practices. *Journal of Internet Services and Applications*, *8*(1), 9. https://doi.org/10.1186/s13174-017-0060-5

Shu, T., & Wu, C. Q. (2017). Energy-efficient mapping of large-scale workflows under deadline constraints in big data computing systems. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2017.07.050

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System (pp. 1–10). IEEE. https://doi.org/10.1109/MSST.2010.5496972

Song, J., He, H., Wang, Z., Yu, G., & Pierson, J.-M. (2016). Modulo Based Data Placement Algorithm for Energy Consumption Optimization of MapReduce System. *Journal of Grid Computing*, 1–16. https://doi.org/10.1007/s10723-016-9370-2

Tran, X. T., Do, T. V., Rotter, C., & Hwang, D. (2018). A New Data Layout Scheme for Energy-Efficient MapReduce Processing Tasks. *Journal of Grid Computing*, 1–14. https://doi.org/10.1007/s10723-018-9433-7

Vera-Baquero, A., Colomo-Palacios, R., & Molloy, O. (2016). Real-time business activity monitoring and analysis of process performance on big-data domains. *Telematics and Informatics*, *33*(3), 793–807. https://doi.org/10.1016/j.tele.2015.12.005

Wei, Z., & Ren, D. Q. (2014). Review of energy aware big data computing measurements, benchmark methods and performance analysis. In *2014 23rd International Conference on Computer Communication and Networks (ICCCN)* (pp. 1–4). https://doi.org/10.1109/ICCCN.2014.6911835

Wu, J., Guo, S., Li, J., & Zeng, D. (2016). Big Data Meet Green Challenges: Greening Big Data. *IEEE Systems Journal*, *10*(3), 873–887. https://doi.org/10.1109/JSYST.2016.2550538

Wu, W., Lin, W., Hsu, C.-H., & He, L. (2017). Energy-efficient hadoop for big data analytics and computing: A systematic review and research insights. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2017.11.010

Xue, S., Zhang, Y., Xu, X., Xing, G., Xiang, H., & Ji, S. (2017). QET: a QoS-based energy-aware task scheduling method in cloud environment. *Cluster Computing*, *20*(4), 3199–3212. https://doi.org/10.1007/s10586-017-1047-5

Yang, T., Pen, H., Li, W., & Zomaya, A. Y. (2017). An energy-efficient virtual machine placement and route scheduling scheme in data center networks. *Future Generation Computer Systems, 77*, 1–11. https://doi.org/10.1016/j.future.2017.05.047

Zhai, J., Zhang, H., Zhong, X., Li, W., Wang, L., & He, Z. (2016). Energy-Efficient Hadoop Green Scheduler. In *2016 IEEE First International Conference on Data Science in Cyberspace (DSC)* (pp. 335–340). https://doi.org/10.1109/DSC.2016.11

Zheng, Z., Wu, X., Zhang, Y., Lyu, M. R., & Wang, J. (2013). QoS Ranking Prediction for Cloud Services. *IEEE Transactions on Parallel and Distributed Systems*, *24*(6), 1213–1222. https://doi.org/10.1109/TPDS.2012.285

Zikopoulos, P. (2012). *Understanding big data: analytics for enterprise class Hadoop and streaming data*. Retrieved from http://www.books24x7.com/marc.asp?bookid=65470

Zong, Z., Ge, R., & Gu, Q. (2017). Marcher: A Heterogeneous System Supporting Energy-Aware High Performance Computing and Big Data Analytics. *Big Data Research, 8*, 27–38. https://doi.org/10.1016/j.bdr.2017.01.003