Empirical Big Data Research: A Systematic Literature Mapping

L.W.M. Wienhofen[a,*], B.M. Mathisen[a], D. Roman[b]

[a] *SINTEF ICT, PO Box 4760, Sluppen, NO-7465 Trondheim, Norway*
[b] *SINTEF ICT, P.O. Box 124 Blindern, N-0314 Oslo, Norway*

Abstract

Background: Big Data is a relatively new field of research and technology, and literature reports a wide variety of concepts labeled with Big Data. The maturity of a research field can be measured in the number of publications containing empirical results. In this paper we present the current status of empirical research in Big Data. *Method:* We employed a systematic mapping method with which we mapped the collected research according to the labels Variety, Volume and Velocity. In addition, we addressed the application areas of Big Data. *Results:* We found that 151 of the assessed 1778 contributions contain a form of empirical result and can be mapped to one or more of the 3 V's and 59 address an application area. *Conclusions:* The share of publications containing empirical results is well below the average compared to computer science research as a whole. In order to mature the research on Big Data, we recommend applying empirical methods to strengthen the confidence in the reported results. Based on our trend analysis we consider Variety to be the most promising uncharted area in Big Data.

Keywords: Systematic Mapping, Big Data, Empirical, Trend Analysis, Survey

1. Introduction

A sharp increase in the number of publications related to the Big Data field in the past years makes it difficult to get a good overview of the current state-of-the-art. Big Data technology is diverse and can be applied to many areas. Big Data features in many trend reports and academic publications. In order to get an overview of the field, we have performed a systematic mapping study and assessed to which degree empirical results have been reported. In our study empirical results mean that a technology or concept has been tested and evaluated so that the result can be seen as a part of an evidence base. Concepts or technology that are merely referred to and not tested or evaluated are excluded from this study. Generally speaking, Big Data is a collection of large data sets with a great diversity of types so that it becomes difficult to process by using state-of-the-art data processing approaches or traditional data processing platforms [153]. In a 2011 Gartner report [102] Doug Laney explains the concept of Volume, Variety and Velocity in data management. These are known as the 3V's and characterize the concept of Big Data. In addition to these 3 fundamental V's, many other V's have emerged, though these differ per the special feature the author of these publications happen to need.

In 2012, Gartner revised and gave a more detailed definition[1] as: *Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization". More generally, a data set can be called Big Data if it is formidable to perform capture, curation, analysis and visualization on it at the current technologies.*

[*] Corresponding author
Email addresses: leendert.wienhofen@sintef.no (L.W.M. Wienhofen), bjornmagnus.mathisen@sintef.no (B.M. Mathisen), Dumitru.Roman@sintef.no (D. Roman)
URL: http://www.sintef.no/home/Information-and-Communication-Technology-ICT/ (L.W.M. Wienhofen)

[1] http://www.gartner.com/resId=2057415

NIST [54] defines Big Data as: *Big data consists of advanced techniques that harness independent resources for building scalable data systems when the characteristics of the datasets require new architectures for efficient storage, manipulation, and analysis.*

All agree to the fact that Big Data needs to be big, and in order to be assessed as Big Data, one needs to address at least one of the aspects of volume, velocity or variety. However, when one looks into the literature, one finds quite quickly that publications that through their title, keywords or abstract give the impression to deal with Big Data in fact do not address these aspects. Sjøberg et al. [165] state that empirical research seeks to explore, describe, predict, and explain natural, social, or cognitive phenomena by using evidence based on observation or experience. It involves obtaining and interpreting evidence by, e.g., experimentation, systematic observation, interviews or surveys, or by the careful examination of documents or artifacts. Work done in an empirical manner can be used as an evidence base for further research. In order to separate the sheep from the wool, we committed a systematic mapping study taking into account only publications that provide empirical results or address 3 V aspects of Big Data.

1.1. Study approach and contribution

During our mapping of the Big Data literature we found no systematic review of empirical work carried out in the field of Big Data. We did identify related studies and describe these in Section 4.1. In order to create an overview of the areas that are addressed, this paper describes how we carried out a systematic literature mapping with a method similar to [124] to map existing Big Data literature with a form of empirical evidence to the three V's of Big Data as well as application areas. The method is described in detail in Section 2. We chose to limit the mapping to the 3 V's as these are the fundamental issues for Big Data. Many other V-terms have been defined later though none of these are used consistently, which limits the mapping possibilities.

The main contributions of our study are;

- Systematic mapping of the findings of empirical Big Data Studies
- Identification of Big Data studies containing empirical evidence
- Overview of application areas of empirical Big Data Studies
- Trend analysis of empirical Big Data Research
- Identification of and discussion about related surveys

A summary of some of our conclusions:

- The number of reports on Big Data is rising, both empirical and non-empirical
- Roughly 10 percent of the contributions on Big Data include empirical results
- Application areas have been getting more attention over time
- We recommend applying empirical methods to strengthen the confidence in the reported results
- Based on our trend analysis we consider Variety to be the most promising uncharted area in Big Data

1.2. Structure of this paper

The paper is structured as follows. Section 2 describes the general method employed in the work presented in this paper as well as our specific implementation of this method. The results of the different stages of this work is presented in Section 3. Section 4 presents an analysis of the results and Section 5 discusses the limitations of our study. Finally, Section 6 describes our conclusion and our recommendations for further research.

2. The systematic mapping process

The systematic mapping process is an iterative process where each step builds upon the previous. The process starts with a research question and ends with a systematic map, see Figure 1. We have based our systematic mapping procedure on Kai Petersen's and Robert Feldt's work [152]. In this section we first highlight the step (in a text box) in the systematic mapping process as described by Petersen and Feldt, each step in the process description is then followed by a description of how we implemented this step in the process.

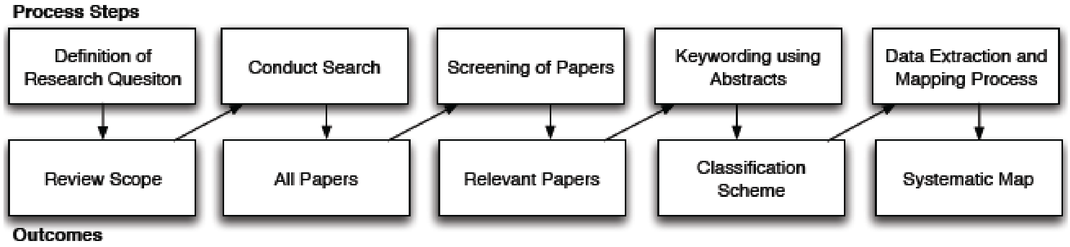

Figure 1: The stages of the systematic mapping process [152].

2.1. Definition of Research Questions (Research Scope)

> The main goal of a systematic mapping study is to provide an overview of a research area, and identify the quantity and type of research and results available within it. Often one wants to map the frequencies of publication over time to see trends. A secondary goal can be to identify the forums in which research in the area has been published.

We have defined the following research questions:

Research Question 1: *Have mapping studies with similar goals to ours been carried out?*

Research Question 2: *What is the share of studies that ground their results with empirical methods?*

Research Question 3: *How are studies that provide empirical results grouped according to the "three Vs"? And what is the distribution of these different groups?*

Research Question 4: *What are the application areas of Big Data and how are they distributed?*

Research Question 5: *Which journals are most prominent?*

Research Question 6: *Can we identify any trends within Empirical Big Data Research?*

2.2. Conduct Search

> The primary studies are identified by using search strings on scientific databases or browsing manually through relevant conference proceedings or journal publications. A good way to create the search string is to structure them in terms of population, intervention, comparison, and outcome [96]. The structure should of course be driven by the research questions. Keywords for the search string can be taken from each aspect of the structure. For example, the outcome of a study (e.g., accuracy of an estimation method) could lead to key words like "case study"

or "experiment" which are research approaches to determine this accuracy.

In this study we used Elsevier's Scopus[2] for our search. Scopus delivers the most comprehensive overview of the world's research output in the fields of science, technology, medicine, social sciences and arts and humanities. It claims to be the largest abstract and citation database of peer-reviewed literature and within our domain it is a valid choice. Test searches done within other databases all returned subsets of the result from Scopus (e.g. the results from "(Big and Data) and (PublishedAs:journal) and (Keywords:Big AND Keywords:Data)" limited to 2014 and earlier in ACM digital library was all part of the Scopus results.)

We searched for "Big Data" in the title, abstract and keywords. We included only papers that are accepted in journals, or in press for journals as well as reviews. This resulted in the following search string:

```
TITLE-ABS-KEY("Big Data") AND DOCTYPE(ar OR re) AND PUBYEAR < 2015
 AND (LIMIT-TO(LANGUAGE,"English"))
   AND (LIMIT-TO(SRCTYPE,"j") OR LIMIT-TO(SRCTYPE,"k"))
```

The string above is defined by the Scopus search query language which can be accessed at `http://www.scopus.com/search/form.url?display=advanced` where `DOCTYPE(ar OR re)` limits to article or review, `PUBYEAR < 2015` limits to publication from before 2015, `(LIMIT-TO(LANGUAGE,"English"))` limits to publications with English as the source language and `LIMIT-TO(SRCTYPE,"j") OR LIMIT-TO(SRCTYPE,"k")` limits to journals and book series.

At the time of the query this also resulted in three publications [23, 74, 195] dated 2015 in direct conflict with the query parameters. This may be because of some journal predating an article (indexing it when it is accepted, but before it is actually published). We have included these papers in the data for completeness. These publications were later excluded in our selection process and thus will not be included in any of the analysis except from the graphs presenting the total number of publication and any calculation or analysis that depends on the total number of publications.

2.3. Screening of Papers for Inclusion and Exclusion (Relevant Papers)

Inclusion and exclusion criteria are used to exclude studies that are not relevant to answer the research questions. The criteria show that the research questions influenced the inclusion and exclusion criteria. It is useful to prototype inclusion and exclusion parameters with a limited set of papers.

Rather than having specific inclusion criteria other than the selection by the search query described above, our method includes every paper from the base corpus until excluded. Exclusion criteria are used to exclude studies that are not relevant to answer the research questions. One may however regard the inverse of criterion 3 and 4 as inclusion criteria. See Table 1 for the criteria.

Given the inclusion and exclusion criteria used in the study described in the process description, we defined criteria suitable for our data-set. Out of the full dataset, we randomly chose 100 papers to test the inclusion and exclusion parameters, prior to reading the full set of papers. The random selection was based on numbering in Endnote, we simply took approximately every 15th paper and included it in our random set. The first exclusion criteria is "no abstract" (some of the results appeared to be short-papers in magazines, and these typically do not include abstracts), as, if there is no abstract, we simply cannot see whether the publication is relevant or not. The second exclusion criteria is "Source language other than English". Some abstracts were written in a way that we typically see when a machine translation is applied, which caused doubt regarding the actual content of the contribution. Upon further investigation of the meta-data, we noticed

[2] `http://www.scopus.com`

that some contributions do not have English as a source language. This will make us unable to do further investigation when needed and therefore we chose to exclude these articles. Once the contribution passed the first two criteria, we looked into the content. Only clear contributions to Big Data are of interest for this mapping study and therefore we exclude (criterion 3) abstracts that do not clearly define contribution of work, as well as abstracts (criterion 4) that are clearly not related to the modern term Big Data (e.g. [151] talks about “Big Data reduction”, however it is clearly not related to the modern term “Big Data”). Publications with very small data sets that claim that the solution will work on a huge dataset without having a convincing strategy for this are also excluded by criterion 4. See also Table 1.

Exclusion criteria number	Criteria
1	No abstract.
2	Source language other than English
3	Abstract does not clearly define contribution of work.
4	Clearly not related to Big Data.

Table 1: Exclusion criteria

2.4. Keywording of Abstracts (Classification Scheme)

Keywording is a way to reduce the time needed in developing the classification scheme and ensuring that the scheme takes the existing studies into account.

We adopted the systematic process for classification from [152]. However, instead of searching for keywords to base the cluster map on, in our case, the keywords were already defined by Laney [102] as explained in the introduction. In addition to the 3 V’s we also defined ”application area” as a keyword to map to. As for the mapping to empirical keywords, we extracted a list of empirical method keywords, which has been compiled in a matrix showing co-occurrences in Table 3.

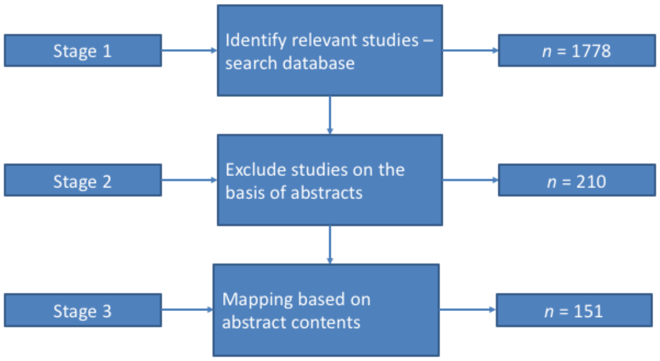
Stage 1
Identify relevant studies – search database
n = 1778
Stage 2
Exclude studies on the basis of abstracts
n = 210
Stage 3
Mapping based on abstract contents
n = 151

Figure 2: The three stages of the systematic mapping.

2.5. Data Extraction and Mapping of Studies (Systematic Map)

> When having the classification scheme in place, the relevant articles are sorted into the scheme, i.e., the actual data extraction takes place. The classification scheme evolves while doing the data extraction, like adding new categories or merging and splitting existing categories. A scheme, for example in Excel, should be used to document the data extraction process. The table should contain each category of the classification scheme. When the reviewers enter the data of a paper into the scheme, they provide a short rationale why the paper should be in a certain category (for example, why the paper applied evaluation research). From the final table, the frequencies of publications in each category can be calculated.
> Mapping data in graphs is a useful aid for the reader to understand the analysis. Visualization alternatives could be found in statistics, HCI and information visualization fields.

We began by categorizing all articles based on their abstracts into four categories. We determine whether the article in question have contributed to the Big Data field itself in term of either volume, variety or velocity. If the article is simply applying one or more Big Data techniques in a case, we identified whether this is a Big Data experiment that has contributed to the field itself by proving that “doing X is possible with these techniques” to a reasonable degree or if this has little effect on the field. In addition, we categorized the contribution according to the empirical methods used:

1. Volume: Describes improvements and progress within technologies and methods for handling increases to the volume of data.
2. Variety: Describes improvements and progress within technologies for handling variety of data.
3. Velocity: Describes improvements and progress within technologies for coping with the speed of incoming data.
4. Application area: Describes Big Data technology different application areas; not innovating through new Big Data technology but through applying Big Data Technology to new areas.

3. Findings from Data

This section describes the outcomes of the process steps described in Section 2; The results chapter map directly to the stages of the systematic mapping process described in Figure 1.

The following section is structured as follows. Subsection 3.1 describes the result of “Definition of research question”. Subsection 3.2 describes the result of “Conduct Search”. Subsection 3.3 describes the result of “Screening of papers”. Subsection 3.4 describes the result of “Keywording using abstracts”. Subsection 3.5 describes the result of “Data Extraction and Mapping Process”

The method used can be summarized in three main stages, as shown in Figure 2.

The outcome of stage 3 is described in Section 3.5.

3.1. Review scope

The goal of this study is to understand the state-of-the-art within the field of Big Data. We aim to identify past and current trends. A secondary goal is to identify the forums that publishes research in the field of study. Our research questions reflects these goals.

3.2. All papers

The query as described in Subsection 2.2 was sent to Elsevier’s SCOPUS February 12th 2015, and resulted in 1778 publications. In Figure 4 we have listed the distribution of publications per year.

3.3. Relevant papers

In Figure 3 we show the number of included papers per year. We also present which of the journals were most prominent in Table 2.

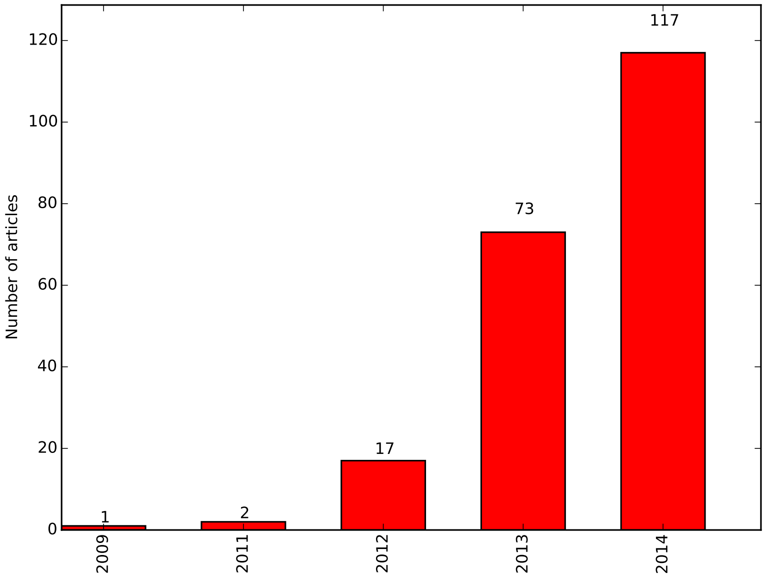

Figure 3: Distribution of the included journal articles according to published year.

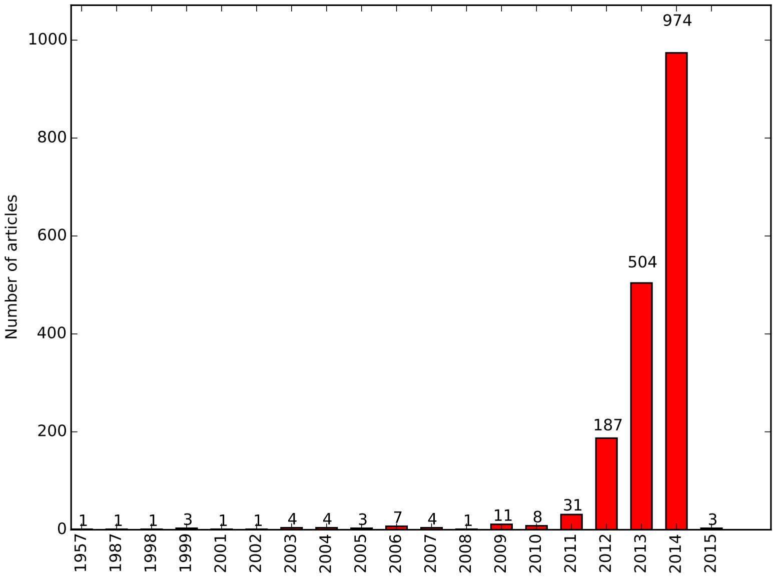

Figure 4: Distribution of the journal articles that contains “Big Data” or “Big Datum” in title, abstracts or keywords according to published year. This includes old uses of the term “Big Data” as can be seen by the publication dated 1957.

Publication source	No.
Lecture Notes in Computer Science	13
Proceedings of the VLDB Endowment	8
Future Generation Computer Systems	7
IEEE Transactions on Emerging Topics in Computing	6
Distributed and Parallel Databases	5
IEEE Network	4
Expert Systems with Applications	4
IEEE Transactions on Knowledge and Data Engineering	4
Journal of Supercomputing	4
Knowledge and Information Systems	4
International Journal of Distributed Sensor Networks	3
Big Data Research	3
PLoS ONE	3
International Journal of Multimedia and Ubiquitous Engineering	3
International Journal of Communication Systems	3
Parallel Computing	3
Journal of Internet Technology	3
Cluster Computing	3
IEEE Transactions on Parallel and Distributed Systems	3
Neural Networks	2
International Journal of Approximate Reasoning	2
NTT Technical Review	2
Studies in Computational Intelligence	2
Computing	2
Journal of Parallel and Distributed Computing	2
ACM Transactions on Knowledge Discovery from Data	2
Pattern Recognition Letters	2
Performance Evaluation	2
Computer Science and Information Systems	2
Tsinghua Science and Technology	2
International Review on Computers and Software	2
International Journal of Business Process Integration and Management	2
Concurrency Computation Practice and Experience	2
Machine Learning	2

Table 2: Publication sources (labeled as journals by Scopus) by number of included publications.

3.4. Classification scheme

First we produced our primary corpus by applying the exclusion criteria (see Table 1) to the initial population of papers described in the previous section. When reading the title and abstract we first checked if the paper was affected by any of our exclusion parameters. After this check was passed, keywording was applied in the sense that main contributions were highlighted in the meta-data. This process is outlined in Figure 5. Some articles turned out to be application area descriptions rather than contributions to any of the V's. These were classified accordingly.

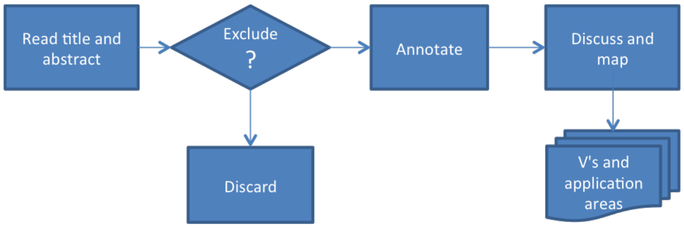

Figure 5: A flowchart describing the classification scheme applied in this review process.

3.5. Systematic map

Through our classification we mapped the publications onto four categories Volume, Velocity, Variety and Application area according to what they contributed with in the field of Big Data. Table 4 shows the mapping results according to publication year. The publications are also mapped onto a Venn diagram in Figure 6, showing the number of publications in each V and their intersections. The numbers in Figure 6 correspond to Table 6 which also includes the references to the publications. Application areas are listed in Table 10.

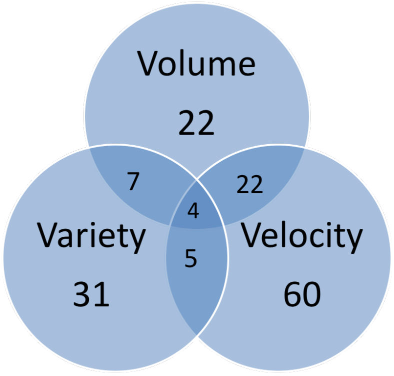

Figure 6: Venn diagram of the mapped studies.

3.6. Empirical Methods

In Table 3 we provide an overview of the means that were identified as methods for being evaluated as being empirical. In the cross matrix we see that some contribution use several methods in order to prove the value of their work.

	Bench.	Case study	Demon.	Eval.	Exp.	Vali.	Impl.	Model	Simul.	Verify
Benchmark	14	0	4	0	8	1	4	0	0	1
Case study	0	6	1	1	0	0	0	0	0	0
Demonstrate	4	1	48	10	32	3	2	3	2	1
Evaluate	0	1	10	35	14	1	4	3	1	1
Experiment	8	0	32	14	108	7	13	4	3	4
Implement	4	0	2	4	13	1	22	1	1	1
Model	0	0	3	3	4	4	1	23	1	0
Simulation	0	0	2	1	3	1	1	1	11	0
Validate	1	0	3	1	7	12	1	4	1	0
Verify	1	0	1	1	4	0	1	0	0	8
Total	32	8	106	70	193	30	49	39	20	16

Table 3: Cross matrix of empirical methods, showing which empirical method keyword was used how many times, and their co-occurrence. The top row is an abbreviated version of the first column.

Reference Year	Variety	Velocity	Volume
2009	1	0	1
2011	0	1	1
2012	1	10	7
2013	19	34	23
2014	26	47	22
Total	47	92	54

Table 4: Three V's according to publishing year.

	2014	2013	2012	2011
Benchmark	8	5	1	0
Case study	4	2	0	0
Demonstrate	25	20	2	1
Evaluate	20	10	4	1
Experiment	66	37	4	0
Implement	14	5	2	1
Model	10	12	1	0
Simulation	5	4	2	0
Validate	7	4	1	0
Verify	4	2	1	1
Total	163	101	18	4

Table 5: Empirical methods according to reference year.

4. Analysis

From the total of 1749 included studies, Figure 3 depicts the distribution of the included studies sorted by publication year. Starting with a total of 3 papers in 2009-2011, we see that in 2012 there was an increase in relevant publications and near fourfold in 2013, this trend continues into 2014. So, we can state there is a clear up-going trend in relevant publications.

Of the 210 included studies, 151 could be mapped onto one or more of the V's, the remaining 59 are papers describing Big Data technologies applied to application areas. In the VENN diagram we chose to exclude to view the application areas and keep the focus on the V's. The most addressed area is velocity with 60 publications, followed by variety with 31 and volume with 22. The combination of volume and velocity is most used with 22 included contributions, whereas volume and variety is mentioned in combination 7 times. Finally, variety and velocity have five included contributions and only four studies [34, 45, 112, 227] mention all three of the main areas of Big Data. From this we can conclude that the most mature areas in terms of published results are Velocity and Volume. We do want to note that many of the contributions mention Hadoop and MapReduce as a basis platform while the focus of content is directed towards velocity and/or variety. This can potentially indicate that the storage is taken for granted when this is used.

V	Publication count	Publications
Variety	31	[2][9][10][14][21] [24][51][60][70][94] [99][100][111][110][108] [122][116][125][126][130] [141][142][146][149][150] [163][173][183][187][215] [231]
Volume	22	[3][12] [19][30][69][81][95] [103][105][120][129][131] [133][166][169][180][192] [203][205][220][221][226]
Velocity	60	[4][5][15][16][27] [26][33][37][42][49] [55][56][57][58][61] [62][64][65][71][75] [76][77][79][80][82] [86][91][92][101][104] [107][113][115][119][134] [140][143][145][148][157] [161][170][181][186][188] [190][196][197][198][199] [202][207][211][212][214] [217][218][225][229][232]
Volume and Velocity	22	[1][7][17][20][25] [29][32][35][39][47] [114][144][177][178][191] [201][206][208][210][213] [219][233]
Velocity And Variety	5	[63][160][168][171][209]
Variety and Volume	7	[43][6][68][123][167] [185][175]
Volume, Variety & Velocity	4	[34][45][112][227]

Table 6: Number of publications classified to that V and the references to those publication.

	total	included	%
before 2009	31	0	0,00
2009	11	1	9,09
2010	8	0	0,00
2011	31	2	6,45
2012	187	17	9,09
2013	504	73	14,48
2014	974	117	12,01
2015	3	0	0,00
sum	1749	210	12,01

Table 7: Total number of journal publications per year.

Table 7 and Figure 8 give an overview of the total number of journal papers per year that we have assessed, as well as the number of included paper. It becomes very clear that the majority of publications do not have empirical findings. These numbers are also presented in Figure 3 (included papers per year) and Figure 4 (total papers per year). The Inclusion percentage is graphed in Figure 9

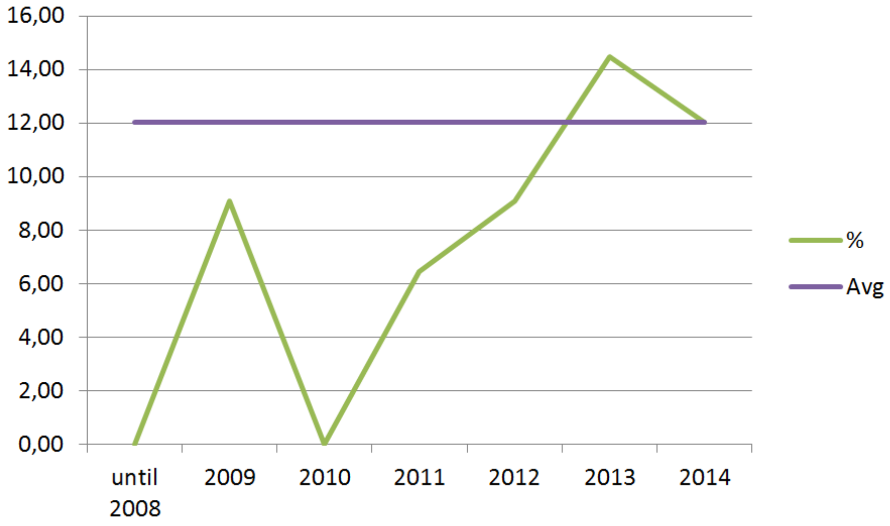

Figure 7: Percentage of empirical (included) studies per year.

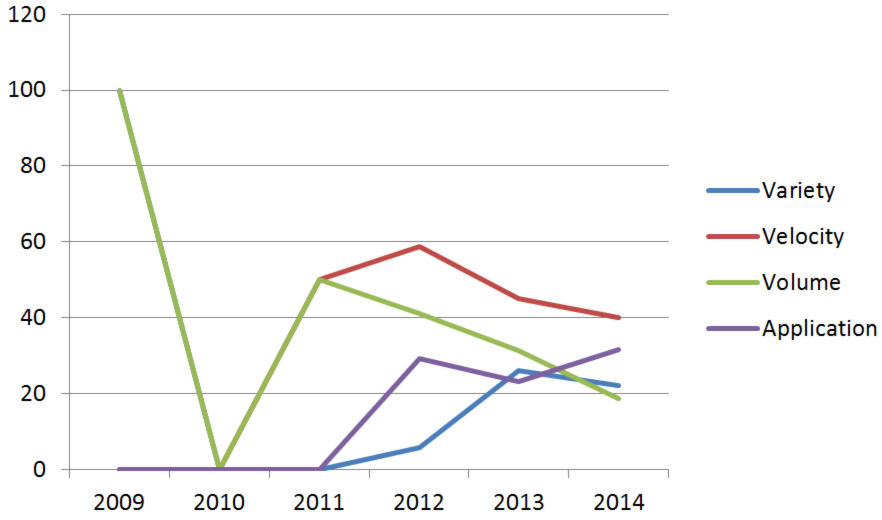

Figure 8: Percentage of empirical (included) studies per year per mapped region: Volume, Velocity, Variety and Application Area.

We have mapped the included studies to 4 categories, Variety, Velocity, Volume and Application area. The latter means that a paper does contribute empirically to Big Data by applying Big Data technology to a field, however, without forwarding the technology itself. Some papers address multiple V's, and if so, are mapped accordingly. Hence, the total number of V's is higher than the total number of included papers.

Table 8 gives an overview of the included papers mapped to according to the V and application area. The column "% change" indicates the change from the year before and is meant to give an indication on whether there is growth. Except for Volume, we see that in the past 3 years there has been an increase for all categories in absolute numbers. Though, measured in percentage growth compared to the previous year we see a decline.

	Variety	% change	Velocity	% change	Volume	% change	Application	% change
2009	1	N/A	0	N/A	1	N/A	0	N/A
2010	-	N/A	-	N/A	-	N/A	-	N/A
2011	0	N/A	1	N/A	1	N/A	0	N/A
2012	1	N/A	10	1000,00	7	700,00	5	N/A
2013	19	1900,00	33	330,00	23	328,57	17	340,00
2014	26	136,84	47	142,42	22	95,65	37	217,65

Table 8: Included publications and their mapping to V and application area. The table also reflects the change over time in percentage.

	Variety	included %	Velocity	included %	Volume	included %	Appl.	included %
2009	1	100,0	0	0,00	1	100,0	0	0,00
2010	-	N/A	-	N/A	-	N/A	-	N/A
2011	0	0,0	1	50,0	1	50,0	0	0,0
2012	1	5,88	10	58,82	7	41,18	5	29,41
2013	19	26,03	33	45,21	23	31,51	17	23,29
2014	26	22,22	47	40,17	22	18,80	37	31,62

Table 9: Overview of the included publications mapped to V's and the applications areas.

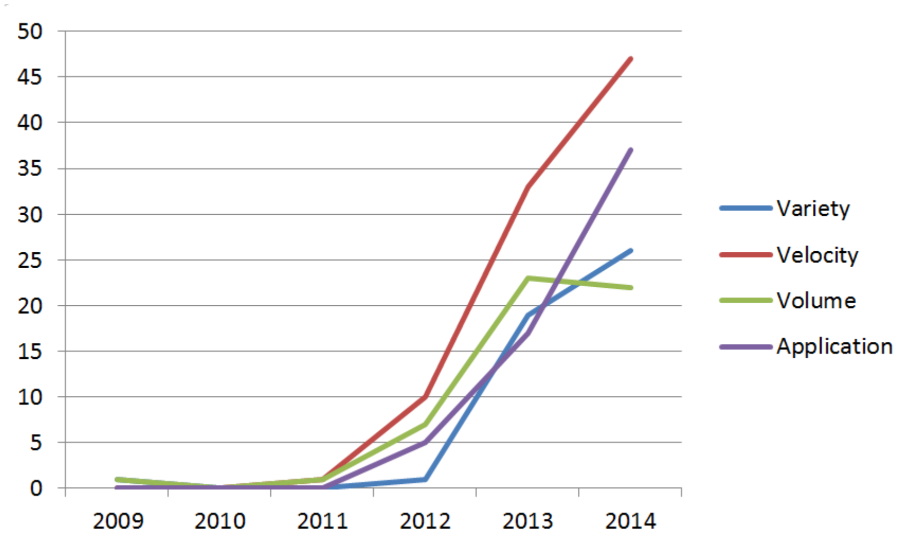

Figure 9: Number of empirical (included) studies per year per mapped region: Volume, Velocity, Variety and Application Area.

Table 9 gives an overview of the included papers mapped according to the V's and application areas. The column "included %" indicates the percentage of included papers. For example, in 2009 we included 1 paper which was mapped on both Variety and Volume (also explaining why the total can be higher than 100% per year). In the past 3 years, for Variety we see that the percentage of included papers has increased quite a bit from 2012 to 2013, with a little decline to 2014. Both Velocity and Volume have declined in the past 3 years. Though, the total number of publications is still increasing (except Volume 2013/2014).

59 of the included studies is not classified as a direct contributor to any of the three V's, rather

the publication describes Big Data technology applied to different domains to such a degree that it can be viewed as a contribution to Big Data as a field.

Application area	Number of publications	Publications
Social (network) analysis	7	[11, 31, 53, 59, 88, 98, 127]
(Cyber) Security and privacy	6	[93, 124, 204, 222, 223, 230]
Visual analytics	5	[50, 73, 121, 132, 207]
Predictive analytics	4	[8, 18, 224, 228]
Intelligent Transport Systems	4	[36, 46, 118, 216]
Search engine/data exploration	3	[97, 106, 156]
Environmental monitoring and management	3	[40, 67, 174]
(Bio)Medical	3	[83, 117, 137]
Text Extraction	3	[72, 87, 155]

Table 10: Publications grouped by application area.

In addition we have studies within Recommendations [138, 162], Cost reduction [28], Image and video classification tasks [41], Stimulation of learning experience [13], Clustering [38], ATC [78], Telecom [89], Cloud [128], Kernel spectral clustering [135, 136] (However these were very close to be classified as applicable to Velocity and Variability), Knowledge provision [139], Smart Grid [164], Analytics [172], Space [176], Criminal investigation [179], Marketing [182], retrieval of learning objects [184], Bibliometrics [200], Service operation [109], recreational studies [193]

We cannot give a conclusive trend analysis based on our study, though as we do have indications, we wanted to see if these coincide with generally available trend reports. Trend reports and predictions are abundant; a quick Google search on "latest trends in Big Data" returns millions of results. At the time of search (on Google), the first hits were:

Gartner Predicts Three Big Data Trends for Business Intelligence[3]:

F1	By 2020, information will be used to reinvent, digitalize or eliminate 80% of business processes and products from a decade earlier.
F2	By 2017, more than 30% of enterprise access to broadly based Big Data will be via intermediary data broker services, serving context to business decisions.
F3	By 2017, more than 20% of customer-facing analytic deployments will provide product tracking information leveraging the IoT

Table 11: Trends within Big Data predicted by Gartner (published by Forbes)

Top Big Data and Analytics Trends for 2015[4]:

[3] `http://www.forbes.com/sites/gartnergroup/2015/02/12/gartner-predicts-three-big-data-trends-for-business-intelligence/`

[4] `http://www.zdnet.com/article/2015-interesting-big-data-and-analytics-trends/`

Z1	More Magic
Z2	Datafication
Z3	Multipolar Analytics
Z4	Fluid Analysis
Z5	Community
Z6	Analytic Ecosystems
Z7	Data Privacy

Table 12: Big Data and Analytics trends predicted in 2015 by ZDnet.

CIO's 5 Big Data Technology Predictions for 2015[5]:

C1	Data Agility Emerges as a Top Focus
C2	Organizations Move from Data Lakes to Processing Data Platforms
C3	Self-Service Big Data Goes Mainstream
C4	Hadoop Vendor Consolidation: New Business Models Evolve
C5	Enterprise Architects Separate the Big Hype from Big Data

Table 13: CIO's Big Data Technology predictions for 2015.

We have enumerated the headings from the trend reports for easier referencing.

We are aware that this is a limited subset of all available trend reports, though these should give at least a general impression and a basis for comparing our trend indication based on the literature study. We have omitted reports that require registration.

Based on our literature study, we have a careful indication that Variety is more on the rise than Velocity and Volume. Also, Big Data technology is becoming more applied.

The latter is reflected in F1, F2 C3, C4, and C5.

Analysis of data is mentioned in F3, Z3, Z4, and Z6. Variety is closely related to analytics.

Volume and Velocity do not seem to be reflected in these reports.

On general terms, we can state that the reports agree that Big Data is becoming more mature and therefore more applied and that analytics is the path to choose if you want to stay in front of the state-of-the art. This is supported by Kambatla et al. [90].

4.1. Related mappings, surveys and reviews

In addition to the above, we also identified studies that did not meet our inclusion criteria; though do provide a contribution in creating an overview of a part of the Big Data field. Below we summarize the type of contribution and their conclusions.

Sakr et al. [159] provide a comprehensive survey for a family of approaches and mechanisms of large-scale data processing mechanisms that have been implemented based on the original idea of the MapReduce framework and are currently gaining a lot of momentum in both research and industrial communities. They also cover a set of introduced systems that have been implemented to provide declarative programming interfaces on top of the MapReduce framework. In addition, they review several large-scale data processing systems that resemble some of the ideas of the MapReduce framework for different purposes and application scenarios

Gorodov and Gubarev [52] have done a review of methods for visualizing data and provided a classification of visualization methods in application to Big Data.

Ruixan [158] presents Bibliometrical Analysis on the Big Data Research in China and summarizes research characteristics in order to study Big Data in-depth development and the future development of Big Data. They also provide reference information for studies related to Library

[5] `http://www.cio.com/article/2862014/big-data/5-big-data-technology-predictions-for-2015.html`

and Information Studies. They conclude that research based on Big Data has taken shape though most of these papers in the theoretical stage of exploration, lack adequate practical support and therefore recommend to intensify efforts based on theory and practice.

Chen and Zhang [153] have done a comprehensive survey of Big Data technologies, techniques, challenges and applications. They offer a close view of Big Data applications opportunities and challenges as well as techniques that is currently adopted and used to solve Big Data problems.

Jeong and Ghani [85] have done a review of semantic technologies for Big Data and conclude that their analysis shows that there is a need to put more effort into proposing new approaches, and that tools must be created that support researchers and practitioners in realizing the true power of semantic computing and solving the crucial issues of Big Data.

Gandomi and Haider [48] present a consolidated description of Big Data by integrating definitions from practitioners and academics. The paper's primary focus is on the analytic methods used for Big Data. A particular distinguishing feature of this paper is its focus on analytics related to unstructured data, which according to these authors constitute 95% of Big Data.

Wang and Krishnan [189] present a review with an objective to provide an overview of the features of clinical Big Data. They describe a few commonly employed computational algorithms, statistical methods, and software tool kits for data manipulation and analysis, and discuss the challenges and limitations in this realm.

Fernández et al. [44] focus on systems for large-scale analytics based on the MapReduce scheme and Hadoop. They identify several libraries and software projects that have been developed for aiding practitioners to address this new programming model. They also analyze the advantages and disadvantages of MapReduce, in contrast to the classical solutions in this field. Finally, they present a number of programming frameworks that have been proposed as an alternative to MapReduce, developed under the premise of solving the shortcomings of this model in certain scenarios and platforms.

Polato et al. [154] have conducted a systematic literature review to assess research contributions to Apache Hadoop. The objective was to identify gaps, providing motivation for new research, and outline collaborations to Apache Hadoop and its ecosystem, classifying and quantifying the main topics addressed in the literature.

Wu and Yamaguchi [194] presents a survey of Big Data in life sciences, Big Data related projects and Semantic Web technologies. The paper helps to understand the role of Semantic Web technologies in the Big Data era and how they provide a promising solution for the Big Data in life sciences.

Kambatla et al. [90] provide an overview of the state-of-the-art and focus on emerging trends to highlight the hardware, software, and application landscape of big-data analytics.

Hashem et al. [66] have assessed the rise of big data in cloud computing. The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology are also discussed. Furthermore, research challenges are investigated, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Lastly, they give a summary of open research issues that require substantial research efforts.

As for ongoing projects, The Byte project (EU FP7) is also investigating the research field of Big Data. And the Big Data Value Association is an initiative with the goal to provide the Big Data Value strategic research agenda (SRIA) and its regular updates, defining and monitoring the metrics of the cPPP and joining the European Commission in the cPPP partnership board.

None of the studies above have thoroughly mapped the existing knowledge against the Big Data V concepts, nor assessed whether the contributions have created empirical results

5. Discussion

There are some limitations to this study. The first limitation is that we used a single source for our search. Scopus[6] claims to be “the largest abstract and citation database of peer-reviewed literature: scientific journals, books and conference proceedings”, making it a valid choice. Scopus also returned a super set of the search results we got from trying the same query on IEEE Xplore, ACM digital library and Compendex. With regards to the mapping processes, one could claim that both researchers should have read all abstracts and discussed all. Instead, we did a pre-mapping in tandem and after that split the work. Each was working for himself, excluding the clear outliers based on the exclusion criteria and including the clear paper to include. In case of even the slightest doubt, we marked the publication and discussed these publications later on. One can also argue that it is a limitation to limit the mapping to the 3V’s as well as application areas and not include the other “V’s” that appear in the papers. We argue that sticking to the original 3 V’s gives a much more concise overview than also including non-standard V’s that emerge. The 3 basic V’s are well defined, whereas others V’s are open for interpretation. Another limitation is the definition of the empirical work. However we do not use the definitions, we just record the words used in abstracts that are also word that describe empirical methods. If an authors claims that they have done an experiment and that the results has been evaluated, we noted this and did not read the full publication in order to investigate if this is really true. We have not assessed the quality of the work carried out in detail, other than noting that the publication is a peer reviewed journal. Finally, we did not find very clear trends in the analyzed data, and the method for comparing the correlation between our result and the non-scientific trend reports can be argued as being weak as we based it on a simple Google search. However, the trends do coincide.

6. Conclusion and Recommendations

Typically a mapping study does not assess quality, though as Big Data is and has been a very ”hot topic” over the past years, the term appears in very many papers including papers that have not contributed to Big Data research. Therefore, we chose to only include papers that have some form of empirical approach in order to eliminate the chance of analyzing papers that are not contributing towards forwarding the evidence base of Big Data research. A total of 210 articles were included and 151 of these have been coded against one or more of the three “main V’s”. In addition, we have an overview of application areas (meaning Big Data technology has been applied, though not contributed to forwarding one of the V’s).

> **Research Question 1:** *Have mapping studies with similar goals to ours been carried out?*

Answer: At the time of search we found [22] and [84] that are labeled as reviews, however they were not systematic. In [147] Park et al. use a systematic approach; the paper presents findings on the social networks of authors in co-authored papers within the Big Data field.

> **Research Question 2:** *What is the share of studies that ground their results with empirical methods?*

Answer: We found that on average a bit less than 10% (for details, refer to Table 7) of the retrieved publications include a form of empirical approach. We also identified the type of empirical method was used. For details see Table 3 and Table 5. In the paper “The Future of Empirical Methods in Software Engineering Research”, Sjøberg et al. [165] state that ”an average of the reviews indicates that about 20% of all papers report empirical studies”. This means that the use of empirical methods in Big Data research is below average.

[6] `www.scopus.com`

> **Research Question 3:** *How are studies that provide empirical results grouped according to the "three Vs"? And what is the distribution of these different groups?*

Answer: We identified papers that could clearly be classified as contributing to the Big Data field within either an application area, or technology within Volume, Velocity or Variety. The analysis that followed revealed that Velocity (90 papers) has received the most attention from researchers, whereas Variety (46 papers) and Volume (54 papers) each have about half of that number. When one looks into the deviation per year, we can see that all V's are still increasing in absolute numbers (with one exception Volume going down from 23 to 22). For more information see section 3 and figure 6.

> **Research Question 4:** *What are the application areas of Big Data and how are they distributed?*

Answer: Big data is within many different application areas, we identified 65 papers describing the use of Big Data technology within an application area. We can see an increase in papers addressing an application area over time. For more details see section 4.

> **Research Question 5:** *Which journals are most prominent?*

Answer: From the studies retrieved by our search, limiting to contributions that are classified as journals by Scopus, we found that **Lecture Notes in Computer Science** is the most prominent channel featured in our selected papers, followed by **Proceedings of the VLDB Endowment** and **Future Generation Computer Systems**, for the full list see table 2.

> **Research Question 6:** *Can we identify any trends within Empirical Big Data Research?*

Answer: Based on our literature study, we have a careful indication that Variety is more on the rise than Velocity and Volume. Also, Big Data technology is becoming more applied. Referring to the analysis in section 4, we can -on general terms- state Big Data is becoming more mature and therefore more applied and that analytics is the path to choose if you want to stay in front of the state-of-the art.

Recommendations: The share of publications containing empirical results is well below the average compared to computer science research as a whole. In order to mature the research on Big Data, we recommend to both use the evidence base of existing empirical studies in Big Data and we recommend applying empirical methods to strengthen the confidence in the reported results. We consider Variety to be the most promising uncharted area in Big Data and recommend focusing research in this direction.

Acknowledgments

We would like to thank Jacqueline Floch, Tore Dybå and Babak Farshchian for their invaluable input and work on this paper. The work presented in this paper is funded by SINTEF ICT.

References

References

[1] C. L. Abad, M. Yuan, C. X. Cai, Y. Lu, N. Roberts, R. H. Campbell, Generating request streams on big data using clustered renewal processes, Performance Evaluation 70 (10) (2013) 704–719.

[2] J. H. Abawajy, A. Kelarev, M. Chowdhury, Large iterative multitier ensemble classifiers for security of big data, IEEE Transactions on Emerging Topics in Computing 2 (3) (2013) 352–363, iEEE Trans. Emerg. Top. Comput.

[3] B. Ahmadi, K. Kersting, M. Mladenov, S. Natarajan, Exploiting symmetries for scaling loopy belief propagation and relational training, Machine Learning 92 (1) (2013) 91–132.

[4] D. Ai-Mei, Research and implementation of support vector machine and its fast algorithm, International Journal of Multimedia and Ubiquitous Engineering 9 (10) (2014) 79–90, int. J. Multimedia Ubiquitous Eng.

[5] N. S. Alghamdi, W. Rahayu, E. Pardede, Semantic-based structural and content indexing for the efficient retrieval of queries over large xml data repositories, Future Generation Computer Systems 37 (2014) 212–231, future Gener Comput Syst.

[6] Ö. G. Ali, K. Yaman, Selecting rows and columns for training support vector regression models with large retail datasets, European Journal of Operational Research 226 (3) (2013) 471–480.

[7] L. Aniello, L. Querzoni, R. Baldoni, High frequency batch-oriented computations over large sliding time windows, Future Generation Computer Systems 43 (2014) 1–11, future Gener Comput Syst.

[8] M. Ballings, D. Van Den Poel, Customer event history for churn prediction: How long is long enough?, Expert Systems with Applications 39 (18) (2012) 13517–13522.

[9] A. Bantouna, G. Poulios, K. Tsagkaris, P. Demestichas, Network load predictions based on big data and the utilization of self-organizing maps, Journal of Network and Systems Management (2013) 1–24.

[10] J. Bendler, S. Wagner, T. Brandt, D. Neumann, Taming uncertainty in big data: Evidence from social media in urban areas, Business and Information Systems Engineering 6 (5) (2014) 279–288, busin. Info. Sys. Eng.

[11] C. A. Bliss, I. M. Kloumann, K. D. Harris, C. M. Danforth, P. S. Dodds, Twitter reciprocal reply networks exhibit assortativity with respect to happiness, Journal of Computational Science 3 (5) (2012) 388–397.

[12] A. Brandau, J. Tolujevs, Modelling and analysis of logistical state data, Transport and Telecommunication 14 (2) (2013) 102–115.

[13] S. Caballé, F. Xhafa, Distributed-based massive processing of activity logs for efficient user modeling in a virtual campus, Cluster Computing (2013) 1–16.

[14] Y. Cai, Q. Li, H. Xie, H. Min, Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy, Neural Networks 58 (2014) 98–110, neural Netw.

[15] S. Campa, M. Danelutto, M. Goli, H. Gonzlez-Vlez, A. M. Popescu, M. Torquati, Parallel patterns for heterogeneous cpu/gpu architectures: Structured parallelism from cluster to cloud, Future Generation Computer Systems 37 (2014) 354–366, future Gener Comput Syst.

[16] A. Cano, S. Ventura, K. J. Cios, Scalable caim discretization on multiple gpus using concurrent kernels, Journal of Supercomputing 69 (1) (2014) 273–292, j Supercomput.

[17] F. Casu, S. Elefante, P. Imperatore, I. Zinno, M. Manunta, C. De Luca, R. Lanari, Sbasdinsar parallel processing for deformation time-series computation, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7 (8) (2014) 3285–3296, iEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.

[18] H. Y. Chang, S. C. Huang, C. C. Lai, A personalized iptv channel-recommendation mechanism based on the mapreduce framework, Journal of Supercomputing 69 (1) (2014) 225–247, j Supercomput.

[19] R. S. Chang, C. S. Liao, K. Z. Fan, C. M. Wu, Dynamic deduplication decision in a hadoop distributed file system, International Journal of Distributed Sensor Networks 2014, int. J. Distrib. Sens. Netw.

[20] G. Chen, K. Chen, D. Jiang, B. C. Ooi, L. Shi, H. T. Vo, S. Wu, E3: An elastic execution engine for scalable data processing, Journal of Information Processing 20 (1) (2012) 65–76.

[21] H. Chen, W. Jiang, C. Li, R. Li, A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm, Mathematical Problems in Engineering 2013, math. Probl. Eng.

[22] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, X. Zhou, Big data challenge: a data management perspective, Frontiers of Computer Science 7 (2) (2013) 157–164.

[23] J. Chen, J. Ma, N. Zhong, Y. Yao, J. Liu, R. Huang, W. Li, Z. Huang, Y. Gao, J. Cao, Waas: Wisdom as a service, IEEE Intelligent Systems 29 (6) (2015) 40–47.

[24] K. Chen, Optimizing star-coordinate visualization models for effective interactive cluster exploration on big data, Intelligent Data Analysis 18 (2) (2014) 117–136, intell. Data Anal.

[25] Y. Chen, Z. Qiao, S. Davis, H. Jiang, K. C. Li, Pipelined multi-gpu mapreduce for big-data processing, in: Studies in Computational Intelligence, 2013.

[26] Y. Chen, X. Qin, H. Bian, J. Chen, Z. Dong, X. Du, Y. Gao, D. Liu, J. Lu, H. Zhang, A study of sql-on-hadoop systems, Lecture Notes in Computer Science 8807 (2014) 154–166, lect. Notes Comput. Sci.

[27] Z. Chen, Y. Lu, N. Xiao, F. Liu, A hybrid memory built by ssd and dram to support in-memory big data analytics, Knowledge and Information Systems 41 (2) (2014) 335–354, knowl. Inf. Systems. Syst.

[28] C. F. Chien, A. C. Diaz, Y. B. Lan, A data mining approach for analyzing semiconductor mes and fdc data to enhance overall usage effectiveness (oue), International Journal of Computational Intelligence Systems 7 (SUPPL.2) (2014) 52–65, int. J. Comput. Intell. Syst.

[29] C. D. Corley, R. M. Farber, W. N. Reynolds, Thought leaders during crises in massive social networks, Statistical Analysis and Data Mining 5 (3) (2012) 205–217.

[30] A. Costan, R. Tudoran, G. Antoniu, G. Brasche, Tomusblobs: Scalable data-intensive processing on azure clouds, Concurrency Computation Practice and Experience.

[31] J. W. Crampton, M. Graham, A. Poorthuis, T. Shelton, M. Stephens, M. W. Wilson, M. Zook, Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb, Cartography and Geographic Information Science 40 (2) (2013) 130–139.

[32] X. Cui, P. Zhu, X. Yang, K. Li, C. Ji, Optimized big data k-means clustering usingmapreduce, Journal of Supercomputing 70 (3) (2014) 1249–1259, j Supercomput.

[33] L. Ding, G. Wang, J. Xin, X. Wang, S. Huang, R. Zhang, Commapreduce: An improvement of mapreduce with lightweight communication mechanisms, Data and Knowledge Engineering.

[34] Z. Ding, Z. Chen, Q. Yang, Iot-svksearch: A real-time multimodal search engine mechanism for the internet of things, International Journal of Communication Systems 27 (6) (2014) 871–897, int J Commun Syst.

[35] Y. Djenouri, H. Drias, Z. Habbas, Bees swarm optimisation using multiple strategies for association rule mining, International Journal of Bio-Inspired Computation 6 (4) (2014) 239–249, int. J. Bio-Inspired Comput.

[36] C. Dobre, F. Xhafa, Intelligent services for big data science, Future Generation Computer Systems 37 (2014) 267–281, future Gener Comput Syst.

[37] S. Ewen, K. Tzoumas, M. Kaufmann, V. Mark, Spinning fast iterative data flows, Proceedings of the VLDB Endowment 5 (11) (2012) 1268–1279.

[38] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, A. Bouras, A survey of clustering algorithms for big data: Taxonomy and empirical analysis, IEEE Transactions on Emerging Topics in Computing 2 (3) (2013) 267–279, iEEE Trans. Emerg. Top. Comput.

[39] W. Fan, F. Geerts, F. Neven, Making queries tractable on big data with preprocessing: (through the eyes of complexity theory), Proceedings of the VLDB Endowment 6 (9) (2013) 685–696.

[40] S. Fang, L. D. Xu, Y. Zhu, J. Ahati, H. Pei, J. Yan, Z. Liu, An integrated system for regional environmental monitoring and management based on internet of things, IEEE Transactions on Industrial Informatics 10 (2) (2014) 1596–1605, iEEE Trans. Ind. Inf.

[41] F. A. Faria, J. A. dos Santos, A. Rocha, R. d. S. Torres, A framework for selection and fusion of pattern classifiers in multimedia recognition, Pattern Recognition Letters.

[42] K. Farrahi, D. Gatica-Perez, A probabilistic approach to mining mobile phone data sequences, Personal and Ubiquitous Computing (2013) 1–16.

[43] Z. Feng, X. Hui-Feng, X. Dong-Sheng, Z. Yong-Heng, Y. Fei, Big data cleaning algorithms in cloud computing, International Journal of Online Engineering 9 (3) (2013) 77–81.

[44] A. Fernández, S. del Río, V. López, A. Bawakid, M. J. del Jesus, J. M. Benítez, F. Herrera, Big data with cloud computing: An insight on the computing environment, mapreduce, and programming frameworks 4 (2014) 380–409–.

[45] P. Ferrera, I. De Prado, E. Palacios, J. L. Fernandez-Marquez, G. Di MarzoSerugendo, Tuple mapreduce and pangool: an associated implementation, Knowledge and Information Systems 41 (2) (2013) 531–557, knowl. Inf. Systems. Syst.

[46] J. Fiosina, M. Fiosins, J. P. Müller, Big data processing and mining for next generation intelligent transportation systems, Jurnal Teknologi (Sciences and Engineering) 63 (3) (2013) 23–38.

[47] A. Floratou, N. Teletia, D. J. DeWitt, J. M. Patel, D. Zhang, Can the elephants handle the nosql onslaught?, Proceedings of the VLDB Endowment 5 (12) (2012) 1712–1723.

[48] A. Gandomi, M. Haider, Beyond the hype: Big data concepts, methods, and analytics 35 (2014) 137–144–.

[49] W. Ge, Y. Huang, D. Zhao, S. Luo, C. Yuan, W. Zhou, Y. Tang, J. Zhou, Cinhba: A secondary index with hotscore caching policy on key-value data store, Lecture Notes in Computer Science 8933 (2014) 602–615, lect. Notes Comput. Sci.

[50] E. Glatz, S. Mavromatidis, B. Ager, X. Dimitropoulos, Visualizing big network traffic data using frequent pattern mining and hypergraphs, Computing (2013) 1–12.

[51] S. Gong, H. Liu, Constructing decision trees for unstructured data, Lecture Notes in Computer Science 8933 (2014) 475–487, lect. Notes Comput. Sci.

[52] E. Y. Gorodov, V. V. Gubarev, Analytical review of data visualization methods in application to big data (2013) –.

[53] A. Gross, D. Murthy, Modeling virtual organizations with latent dirichlet allocation: A case for natural language processing, Neural Networks 58 (2014) 29–37, neural Netw.

[54] N. B. D. W. Group, Nist big data definitions and taxonomies (August 2013).

[55] J. Gu, S. Peng, X. S. Wang, W. Rao, M. Yang, Y. Cao, Cost-based join algorithm selection in hadoop, Lecture Notes in Computer Science 8787 (2014) 246–261, lect. Notes Comput. Sci.

[56] L. Gu, D. Zeng, P. Li, S. Guo, Cost minimization for big data processing in geo-distributed data centers, IEEE Transactions on Emerging Topics in Computing 2 (3) (2013) 314–323, iEEE Trans. Emerg. Top. Comput.

[57] R. Gu, X. Yang, J. Yan, Y. Sun, B. Wang, C. Yuan, Y. Huang, Shadoop: Improving mapreduce performance by optimizing job execution mechanism in hadoop clusters, Journal of Parallel and Distributed Computing 74 (3) (2014) 2166–2179, j. Parallel Distrib. Comput.

[58] Y. Gu, Z. Yang, G. Xu, M. Nakano, M. Toyoda, M. Kitsuregawa, Exploration on efficient similar sentences extraction, World Wide Web 17 (4) (2014) 595–626, world Wide Web.

[59] J. Guan, S. Yao, C. Xu, H. Zhang, Design and implementation of network user behaviors analysis based on hadoop for big data, Communications in Computer and Information Science 490 (2014) 44–55, commun. Comput. Info. Sci.

[60] H. Guang-Ming, C. Nan-Ya, An algorithm in quality inspection of large marine data based on block-nested-loops, BioTechnology: An Indian Journal 8 (2) (2013) 233–237, biotechnol. An Indian J.

[61] S. Gugnani, D. Khanolkar, T. Bihany, N. Khadilkar, Rule based classification on a multi node scalable hadoop cluster, Lecture Notes in Computer Science 8729 (2014) 174–183, lect. Notes Comput. Sci.

[62] C. Guo, W. Luk, Pipelined hac estimation engines for multivariate time series, Journal of Signal Processing Systems 77 (1-2) (2014) 117–129, j. Signal Process Syst.

[63] T. Guo, T. G. Papaioannou, K. Aberer, Efficient indexing and query processing of model-view sensor data in the cloud, Big Data Research 1 (2014) 52–65, big. Data Res.

[64] X. Han, J. Li, D. Yang, J. Wang, Efficient skyline computation on big data, IEEE Transactions on Knowledge and Data Engineering 25 (11) (2013) 2521–2535.

[65] S. Hasan, S. M. Shamsuddin, N. Lopes, Machine learning big data framework and analytics for big data problems, International Journal of Advances in Soft Computing and its Applications 6 (2), int. J. Adv. Soft Comput. Appl.

[66] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, The rise of big data on cloud computing: Review and open research issues, Information Systems 47 (2015) 98 – 115.

[67] M. M. Hassan, B. Song, M. Shamim Hossain, A. Alam, Efficient resource scheduling for big data processing in cloud platform, Lecture Notes in Computer Science 8729 (2014) 51–63, lect. Notes Comput. Sci.

[68] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, M. Palaniswami, Fuzzy c-means algorithms for very large data, IEEE Transactions on Fuzzy Systems 20 (6) (2012) 1130–1146.

[69] B. He, Y. Li, H. Huang, H. Tang, Spatialtemporal compression and recovery in a wireless sensor network in an underground tunnel environment, Knowledge and Information Systems 41 (2) (2014) 449–465, knowl. Inf. Systems. Syst.

[70] T. He, S. Zhang, J. Xin, P. Zhao, J. Wu, X. Xian, C. Li, Z. Cui, An active learning approach with uncertainty, representativeness, and diversity, Scientific World Journal 2014, sci. World J.

[71] H. Herodotou, S. Babu, Profiling, what-if analysis, and costbased optimization of mapreduce programs, Proceedings of the VLDB Endowment 4 (11) (2011) 1111–1122.

[72] T. H. Hong, C. H. Yun, J. W. Park, H. G. Lee, H. S. Jung, Y. W. Lee, Big data processing with mapreduce for e-book, International Journal of Multimedia and Ubiquitous Engineering 8 (1) (2013) 151–162.

[73] G. Hrovat, G. Stiglic, P. Kokol, M. Ojsteršek, Contrasting temporal trend discovery for large healthcare databases, Computer Methods and Programs in Biomedicine.

[74] C. Hsu, B. Zeng, M. Zhang, A novel group key transfer for big data security, Applied Mathematics and Computation 249 (2015) 436–443.

[75] J. Hu, F. L. Lewis, O. P. Gan, G. H. Phua, L. L. Aw, Discrete-event shop-floor monitoring system in rfid-enabled manufacturing, IEEE Transactions on Industrial Electronics 61 (12) (2014) 7083–7091, iEEE Trans Ind Electron.

[76] F. Hueske, M. Peters, M. J. Sax, A. Rheinländer, R. Bergmann, A. Krettek, K. Tzoumas, Opening the black boxes in data flow optimization, Proceedings of the VLDB Endowment 5 (11) (2012) 1256–1267.

[77] T. Hunter, T. Das, M. Zaharia, P. Abbeel, A. M. Bayen, Large-scale estimation in cyber-physical systems using streaming data: A case study with arterial traffic estimation, IEEE Transactions on Automation Science and Engineering 10 (4) (2013) 884–898.

[78] C. Hurter, S. Conversy, D. Gianazza, A. C. Telea, Interactive image-based information visualization for aircraft trajectory analysis, Transportation Research Part C: Emerging Technologies 47 (P2) (2014) 207–227, transp. Res. Part C Emerg. Technol.

[79] S. Ibrahim, H. Jin, L. Lu, B. He, G. Antoniu, S. Wu, Handling partitioning skew in mapreduce using leen, Peer-to-Peer Networking and Applications (2013) 1–16.

[80] A. S. Iwashita, J. P. Papa, A. N. Souza, A. X. Falco, R. A. Lotufo, V. M. Oliveira, V. H. C. De Albuquerque, J. M. R. S. Tavares, A path- and label-cost propagation approach to speedup the training of the optimum-path forest classifier, Pattern Recognition Letters 40 (1) (2014) 121–127, pattern Recogn. Lett.

[81] C. Jardak, P. Mhnen, J. Riihijrvi, Spatial big data and wireless networks: Experiences, applications, and research challenges, IEEE Network 28 (4) (2014) 26–31, iEEE Network.

[82] C. Jayalath, J. Stephen, P. Eugster, From the cloud to the atmosphere: Running mapreduce across data centers, IEEE Transactions on Computers 63 (1) (2014) 74–87, iEEE Trans Comput.

[83] C. P. Jayapandian, C. H. Chen, A. Bozorgi, S. D. Lhatoo, G. Q. Zhang, S. S. Sahoo, Cloudwave: distributed processing of "big data" from electrophysiological recordings for epilepsy clinical research using hadoop, AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2013 (2013) 691–700, aMIA Annu Symp Proc.

[84] K. Jee, G.-H. Kim, Potentiality of big data in the medical sector: focus on how to reshape the healthcare system, Healthcare informatics research 19 (2) (2013) 79–85.

[85] S. R. Jeong, I. Ghani, Semantic computing for big data: Approaches, tools, and emerging directions (2011-2014) (2014).

[86] C. Ji, Z. Li, W. Qu, Y. Xu, Y. Li, Scalable nearest neighbor query processing based on inverted grid index, Journal of Network and Computer Applications 44 (2014) 172–182, j Network Comput Appl.

[87] Y. K. Ji, Y. I. Kim, S. Park, Big data summarization using semantic feture for iot on cloud, Contemporary Engineering Sciences 7 (21-24) (2014) 1095–1103, contemporary Engineering Sciences.

[88] C. Jiang, Y. Chen, K. J. R. Liu, Graphical evolutionary game for information diffusion over social networks, IEEE Journal on Selected Topics in Signal Processing 8 (4) (2014) 524–536, iEEE J. Sel. Top. Sign. Proces.

[89] L. Jun, L. Tingting, C. Gang, Y. Hua, L. Zhenming, Mining and modelling the dynamic patterns of service providers in cellular data network based on big data analysis, China Communications 10 (12) (2013) 25–36, china Commun.

[90] K. Kambatla, G. Kollias, V. Kumar, A. Grama, Trends in big data analytics 74 (2014) 2561–2573–.

[91] K. Kc, V. W. Freeh, Tuning hadoop map slot value using cpu metric, Lecture Notes in Computer Science 8807 (2014) 141–153, lect. Notes Comput. Sci.

[92] V. Khakhutskyy, M. Hegland, Parallel fitting of additive models for regression, Lecture Notes in Computer Science 8736 (2014) 243–254, lect. Notes Comput. Sci.

[93] M. K. Kim, H. J. La, S. D. Kim, A software framework for efficient iot contexts acquisition and big data analytics, Journal of Internet Technology 15 (6) (2014) 939–947, j. Internet Technol.

[94] W. Kim, H. Kim, Y. Kim, Dataconnector: A data processing framework integrating hadoop and a grid middleware ogsa-dai for cloud environment, Information (Japan) 16 (1 B) (2013) 801–806.

[95] Y. Kim, K. Shim, M. S. Kim, J. Sup Lee, Dbcure-mr: An efficient density-based clustering algorithm for large data using mapreduce, Information Systems 42 (2014) 15–35, inf. Syst.

[96] B. A. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering.

[97] I. Kitsos, K. Magoutis, Y. Tzitzikas, Scalable entity-based summarization of web search results using mapreduce, Distributed and Parallel Databases 32 (3) (2014) 405–446, distrib Parallel Databases.

[98] G. H. Knudsen, D. Kjeldgaard, Online reception analysis: Big data in qualitative marketing research, Research in Consumer Behavior 16 (2014) 217–242, res. Consum. Behav.

[99] L. Kuang, F. Hao, L. T. Yang, M. Lin, C. Luo, G. Min, A tensor-based approach for big data representation and dimensionality reduction, IEEE Transactions on Emerging Topics in Computing 2 (3) (2013) 280–291, iEEE Trans. Emerg. Top. Comput.

[100] O. Kwon, J. M. Sim, Effects of data set features on the performances of classification algorithms, Expert Systems with Applications 40 (5) (2013) 1847–1857.

[101] W. Lam, L. Liu, S. Prasad, A. Rajaraman, Z. Vacheri, A. H. Doan, Muppet: Map reduce style processing of fast data, Proceedings of the VLDB Endowment 5 (12) (2012) 1814–1825.

[102] D. Laney, 3d data management: Controlling data volume, velocity, and variety, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[103] N. Laptev, K. Zeng, C. Zaniolo, Early accurate results for advanced analytics on mapreduce, Proceedings of the VLDB Endowment 5 (10) (2012) 1028–1039.

[104] D. Lasalle, G. Karypis, Mpi for big data: New tricks for an old dog, Parallel Computing 40 (10) (2014) 754–767, parallel Comput.

[105] D. Lee, J. S. Kim, S. Maeng, Large-scale incremental processing with mapreduce, Future Generation Computer Systems.

[106] T. Lee, H. Lee, K. H. Rhee, S. U. Shin, The efficient implementation of distributed indexing with hadoop for digital investigations on big data, Computer Science and Information Systems 11 (3) (2014) 1037–1054, comp. Sci. Info. Sys.

[107] C. Li, J. Chen, C. Jin, R. Zhang, A. Zhou, Mr-tree: An efficient index for mapreduce, International Journal of Communication Systems.

[108] M. X. Li, V. Palchykov, Z. Q. Jiang, K. Kaski, J. Kertsz, S. Miccich, M. Tumminello, W. X. Zhou, R. N Mantegna, Statistically validated mobile communication networks: The evolution of motifs in european and chinese data, New Journal of Physics 16, new J. Phys.

[109] T. Li, R. J. Kauffman, Adaptive learning in service operations, Decision Support Systems 53 (2) (2012) 306–319.

[110] X. Li, Y. Wang, X. Li, X. Wang, J. Yu, Gdps: An efficient approach for skyline queries over distributed uncertain data, Big Data Research 1 (2014) 23–36, big. Data Res.

[111] Z. Li, M. A. Sharaf, L. Sitbon, X. Du, X. Zhou, Core: A context-aware relation extraction method for relation completion, IEEE Transactions on Knowledge and Data Engineering 26 (4) (2014) 836–849, iEEE Trans Knowl Data Eng.

[112] F. Liang, C. Feng, X. Lu, Z. Xu, Performance benefits of datampi: A case study with bigdatabench, Lecture Notes in Computer Science 8807 (2014) 111–123, lect. Notes Comput. Sci.

[113] Y. Liang, Y. Wang, M. Fan, C. Zhang, Y. Zhu, Predoop: Preempting reduce task for job execution accelerations, Lecture Notes in Computer Science 8807 (2014) 167–180, lect. Notes Comput. Sci.

[114] M. Lin, L. Zhang, A. Wierman, J. Tan, Joint optimization of overlapping phases in mapreduce, Performance Evaluation 70 (10) (2013) 720–735.

[115] X. Y. Lin, Y. C. Chung, Master-worker model for mapreduce paradigm on the tile64 many-core platform, Future Generation Computer Systems.

[116] D. Liu, T. Li, J. Zhang, A rough set-based incremental approach for learning knowledge in dynamic incomplete information systems, International Journal of Approximate Reasoning 55 (8) (2014) 1764–1786, int J Approximate Reasoning.

[117] F. Liu, H. Yu, Learning to rank figures within a biomedical article, PLoS ONE 9 (3), pLoS ONE.

[118] J. Liu, F. Liu, N. Ansari, Monitoring and analyzing big traffic data of a large-scale cellular network with hadoop, IEEE Network 28 (4) (2014) 32–39, iEEE Network.

[119] J. Liu, C. H. Xia, N. B. Shroff, X. Zhang, On distributed computation rate optimization for deploying cloud computing programming frameworks, Performance Evaluation Review 40 (4) (2013) 63–72.

[120] K. Liu, G. Xu, J. Yuan, An improved hadoop data load balancing algorithm, Journal of Networks 8 (12) (2013) 2816–2822, j. Netw.

[121] S. Liu, W. Cui, Y. Wu, M. Liu, A survey on information visualization: recent advances and challenges, Visual Computer 30 (12), visual Comput.

[122] W. Liu, L. Wang, M. Yi, Simple-random-sampling-based multiclass text classification algorithm, The Scientific World Journal 2014, sci. World J.

[123] Z. Liu, B. Jiang, J. Heer, Immens: Real-time visual querying of big data, Computer Graphics Forum 32 (3 PART4) (2013) 421–430.

[124] Z. Liu, J. Li, J. Li, C. Jia, J. Yang, K. Yuan, Sql-based fuzzy query mechanism over encrypted database, International Journal of Data Warehousing and Mining 10 (4) (2014) 71–87, int. J. Data Warehouse. Min.

[125] R. K. Lomotey, R. Deters, Analytics-as-a-service framework for terms association mining in unstructured data, International Journal of Business Process Integration and Management 7 (1) (2014) 49–61, int. J. Bus. Process Integr. Manage.

[126] J. Lu, D. Li, Bias correction in a small sample from big data, IEEE Transactions on Knowledge and Data Engineering 25 (11) (2013) 2658–2663.

[127] J. Lu, H. Wang, Variance reduction in large graph sampling, Information Processing and Management 50 (3) (2014) 476–491, inf. Process. Manage.

[128] C. Luo, J. Zhan, Z. Jia, L. Wang, G. Lu, L. Zhang, C. Z. Xu, N. Sun, Cloudrank-d: Benchmarking and ranking cloud computing systems for data processing applications, Frontiers of Computer Science in China 6 (4) (2012) 347–362.

[129] T. Luo, W. Yuan, P. Deng, Y. Zhang, G. Chen, A hybrid system of hadoop and dbms for earthquake precursor application, International Review on Computers and Software 8 (2) (2013) 463–467.

[130] K. Ma, F. Dong, B. Yang, Incremental object matching approach of schema-free data with mapreduce, International Journal of Computers and Applications 36 (2) (2014) 72–77, int J Comput Appl.

[131] Y. Ma, X. Meng, Set similarity join on massive probabilistic data using mapreduce, Distributed and Parallel Databases 32 (3) (2014) 447–464, distrib Parallel Databases.

[132] Y. Ma, L. Wang, A. Y. Zomaya, D. Chen, R. Ranjan, Task-tree based large-scale mosaicking for massive remote sensed imageries with dynamic dag scheduling, IEEE Transactions on Parallel and Distributed Systems 25 (8) (2014) 2126–2137, iEEE Trans Parallel Distrib Syst.

[133] R. Maeda, N. Ohta, K. Kuwabara, Mapreduce-based implementation of a rule system, in: Studies in Computational Intelligence, 2014.

[134] A. Majkowska, D. Zydek, L. Koszałka, Task allocation in distributed mesh-connected machine learning system: Simplified busy list algorithm with q-learning based queuing, in: Advances in Intelligent Systems and Computing, 2013.

[135] R. Mall, R. Langone, J. A. K. Suykens, Kernel spectral clustering for big data networks, Entropy 15 (5) (2013) 1567–1586.

[136] R. Mall, R. Langone, J. A. K. Suykens, Multilevel hierarchical kernel spectral clustering for real-life large scale complex networks, PLoS ONE 9 (6), pLoS ONE.

[137] M. Mancini, Exploiting big data for improving healthcare services, Journal of E-Learning and Knowledge Society 10 (2) (2014) 23–33, j. E-learn. Knowl. Soc..

[138] S. Meng, W. Dou, X. Zhang, J. Chen, Kasr: A keyword-aware service recommendation method on mapreduce for big data applications, IEEE Transactions on Parallel and Distributed Systems 25 (12) (2014) 3221–3231, iEEE Trans Parallel Distrib Syst.

[139] D. F. Millie, G. R. Weckman, W. A. Young Ii, J. E. Ivey, D. P. Fries, E. Ardjmand, G. L. Fahnenstiel, Coastal 'big data' and nature-inspired computation: Prediction potentials, uncertainties, and knowledge derivation of neural networks for an algal metric, Estuarine, Coastal and Shelf Science.

[140] H. Mohamed, S. Marchand-Maillet, Mro-mpi: Mapreduce overlapping using mpi and an optimized data exchange policy, Parallel Computing.

[141] M. Muja, D. G. Lowe, Scalable nearest neighbor algorithms for high dimensional data, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (11) (2014) 2227–2240, iEEE Trans Pattern Anal Mach Intell.

[142] F. Narducci, C. Musto, G. Semeraro, P. Lops, M. de Gemmis, Exploiting big data for enhanced representations in content-based recommender systems, in: Lecture Notes in Business Information Processing, 2013.

[143] K. Niu, F. Zhao, S. Zhang, A fast classification algorithm for big data based on knn, Journal of Applied Sciences 13 (12) (2013) 2208–2212.

[144] S. Oda, K. Uenishi, S. Kinoshita, Jubatus: Scalable distributed processing framework for realtime analysis of big data, NTT Technical Review 10 (6).

[145] C. Ordonez, N. Mohanam, C. Garcia-Alvarado, Pca for large data sets with parallel data summarization, Distributed and Parallel Databases 32 (3) (2014) 377–403, distrib Parallel Databases.

[146] A. Papageorgiou, M. Schmidt, J. Song, N. Kami, Efficient filtering processes for machine-to-machine data based on automation modules and data-agnostic algorithms, International Journal of Business Process Integration and Management 7 (1) (2014) 73–86, int. J. Bus. Process Integr. Manage.

[147] H. W. Park, L. Leydesdorff, Decomposing social and semantic networks in emerging "big data" research, Journal of Informetrics 7 (3) (2013) 756–765.

[148] H. W. Park, I. Y. Yeo, J. R. Lee, H. Jang, Study on network architecture of big data center for the efficient control of huge data traffic, Computer Science and Information Systems 11 (3) (2014) 1113–1126, comp. Sci. Info. Sys.

[149] J. Park, H. Kim, Y. S. Jeong, E. Lee, Two-phase grouping-based resource management for big data processing in mobile cloud computing, International Journal of Communication Systems.

[150] J. Peng, G. Seetharaman, W. Fan, A. Varde, Exploiting fisher and fukunaga-koontz transforms in chernoff dimensionality reduction, ACM Transactions on Knowledge Discovery from Data 7 (2), aCM Trans. Knowl. Discov. Data.

[151] E. Perotto, Network operating system, Computer Physics Communications 45 (1-3) (1987) 455–466, export Date: 31 October 2013 Source: Scopus CODEN: CPHCB Language of Original Document: English Correspondence Address: Perotto, E.; CNUCE, Pisa, Italy.

[152] K. Petersen, R. Feldt, Systematic mapping studies in software engineering, in: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, 2008, pp. 71–80.

[153] C. L. Philip Chen, C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data 275 (2014) 314–347–.

[154] I. Polato, R. R, A. Goldman, F. Kon, A comprehensive view of hadoop research - a systematic literature review 46 (2014) 1–25–.

[155] J. Powell, L. Collins, A. Eberhardt, D. Izraelevitz, J. Roman, T. Dufresne, M. Scott, M. Blake, G. Grider, "at scale" author name matching with hadoop/mapreduce, Library Hi Tech News 29 (4) (2012) 6–12.

[156] C. Qin, F. Rusu, Pf-ola: A high-performance framework for parallel online aggregation, Distributed and Parallel Databases 32 (3) (2014) 337–375, distrib Parallel Databases.

[157] S. Richter, J. A. Quian-Ruiz, S. Schuh, J. Dittrich, Towards zero-overhead static and adaptive indexing in hadoop, VLDB Journal 23 (3) (2014) 469–494, vLDB J.

[158] Y. Ruixian, Bibliometrical analysis on the big data research in china 11 (2013) 383–390–.

[159] S. Sakr, A. Liu, A. G. Fayoumi, The family of mapreduce and large-scale data processing systems 46 (2013) –.

[160] E. Sapin, E. Keedwell, A subset-based ant colony optimisation with tournament path selection for high-dimensional problems, Lecture Notes in Computer Science 8790 (2014) 232–247, lect. Notes Comput. Sci.

[161] D. Seo, S. Shin, Y. Kim, H. Jung, S. K. Song, Dynamic hilbert curve-based b+-tree to manage frequently updated data in big data applications, Life Science Journal 11 (10) (2014) 454–461, life Sci. J.

[162] Y. Seo, J. Ahn, Enhancing user-friendliness of the user taste prediction service using mapreduce framework, International Journal of Multimedia and Ubiquitous Engineering 9 (5) (2014) 263–271, int. J. Multimedia Ubiquitous Eng.

[163] T. Silva, M. Jian, Y. Chen, Process analytics approach for r&d project selection, ACM Transactions on Management Information Systems 5 (4), aCM Trans. Manage. Inf. Syst.

[164] Y. Simmhan, S. Aman, A. Kumbhare, R. Liu, S. Stevens, Q. Zhou, V. Prasanna, Cloud-based software platform for big data analytics in smart grids, Computing in Science and Engineering 15 (4) (2013) 38–47, comput. Sci. Eng.

[165] D. I. Sjoberg, T. Dyba, M. Jorgensen, The future of empirical methods in software engineering research, in: 2007 Future of Software Engineering, IEEE Computer Society, 2007, pp. 358–378.

[166] K. Slagter, C. H. Hsu, Y. C. Chung, D. Zhang, An improved partitioning mechanism for optimizing massive data analysis using mapreduce, Journal of Supercomputing (2013) 1–17.

[167] M. Song, H. Yang, S. H. Siadat, M. Pechenizkiy, A comparative study of dimensionality reduction techniques to enhance trace clustering performances, Expert Systems with Applications 40 (9) (2013) 3722–3737.

[168] L. A. Steffene, O. Flauzac, A. S. Charao, P. P. Barcelos, B. Stein, G. Cassales, S. Nesmachnow, J. Rey, M. Cogorno, M. Kirsch-Pinheiro, C. Souveyet, Mapreduce challenges on pervasive grids, Journal of Computer Science 10 (11) (2014) 2194–2210, j. Comput. Sci.

[169] Y. Su, G. Agrawal, J. Woodring, K. Myers, J. Wendelberger, J. Ahrens, Effective and efficient data sampling using bitmap indices, Cluster Computing 17 (4) (2014) 1081–1100, cluster Comput.

[170] M. A. Suchard, S. E. Simpson, I. Zorych, P. Ryan, D. Madigan, Massive parallelization of serial inference algorithms for a complex generalized linear model, ACM Transactions on Modeling and Computer Simulation 23 (1).

[171] H. Sun, J. Weng, G. Yu, R. H. Massawe, A dna-based semantic fusion model for remote sensing data, PLoS ONE 8 (10).

[172] N. Sun, J. G. Morris, J. Xu, X. Zhu, M. Xie, Icare: A framework for big data-based banking customer analytics, IBM Journal of Research and Development 58 (5-6), iBM J. Res. Dev.

[173] M. Tan, I. W. Tsang, L. Wang, Towards ultrahigh dimensional feature selection for big data, Journal of Machine Learning Research 15 (2014) 1371–1429, j. Mach. Learn. Res.

[174] H. Tang, X. Yang, Y. Zhang, Effort at constructing big data sensor networks for monitoring greenhouse gas emission, International Journal of Distributed Sensor Networks 2014, int. J. Distrib. Sens. Netw.

[175] J. Tang, X. Zhao, B. Ge, W. Xiao, H. Shang, Efficiently comparing provenance for knowledge discovery, Journal of Internet Technology 15 (6) (2014) 963–974, j. Internet Technol.

[176] D. Tapiador, W. O'Mullane, A. G. A. Brown, X. Luri, E. Huedo, P. Osuna, A framework for building hypercubes using mapreduce, Computer Physics Communications 185 (5) (2014) 1429–1438, comput Phys Commun.

[177] K. M. Ting, T. Washio, J. R. Wells, F. T. Liu, S. Aryal, Demass: A new density estimator for big data, Knowledge and Information Systems 35 (3) (2013) 493–524.

[178] K. Tretyakov, S. Laur, G. Smant, J. Vilo, P. Prins, Fast probabilistic file fingerprinting for big data, BMC genomics 14 Suppl 2.

[179] M. van Banerveld, N. A. Le-Khac, M. T. Kechadi, Performance evaluation of a natural language processing approach applied in white collar crime investigation, Lecture Notes in Computer Science 8860 (2014) 29–43, lect. Notes Comput. Sci.

[180] D. Venkitaramanan, N. Vijayaraju, K. Velusamy, G. Suresh, M. Divya, Comparison of hadoop multiple node cluster performance over physical and virtual nodes using inverted index data structure for search over wikipedia data set, International Journal of Applied Engineering Research 9 (16) (2014) 3515–3532, int. J. Appl. Eng. Res.

[181] A. Verma, L. Cherkasova, V. S. Kumar, R. H. Campbell, Deadline-based workload management for mapreduce environments: Pieces of the perfromance puzzle, HP Laboratories Technical Report (82).

[182] I. F. Videla-Cavieres, S. A. Ríos, Extending market basket analysis with graph mining techniques: A real case, Expert Systems with Applications.

[183] S. Villa, M. Rossetti, Learning continuous time bayesian network classifiers using mapreduce, Journal of Statistical Software 62 (3) (2014) 1–25, j. Stat. Software.

[184] E. A. Vimal, S. Chandramathi, Learning objects retrieval algorithm using semantic annotation and new matching score, International Review on Computers and Software 8 (12) (2013) 2755–2764, int. Rev. Comput. Softw.

[185] M. Štencl, J. Štastný, Advanced approach to numerical forecasting using artificial neural networks, Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis 57 (6) (2009) 297–304.

[186] J. Wang, Q. Xiao, J. Yin, P. Shang, Draw: A new data-grouping-aware data placement scheme for data intensive applications with interest locality, IEEE Transactions on Magnetics 49 (6) (2013) 2514–2520.

[187] J. Wang, P. Zhao, S. C. H. Hoi, R. Jin, Online feature selection and its applications, IEEE Transactions on Knowledge and Data Engineering 26 (3) (2014) 698–710, iEEE Trans Knowl Data Eng.

[188] L. Wang, L. Feng, J. Zhang, P. Liao, An efficient algorithm of frequent itemsets mining based on mapreduce, Journal of Information and Computational Science 11 (8) (2014) 2809–2816, j. Inf. Comput. Sci.

[189] W. Wang, E. Krishnan, Big data and clinicians: A review on the state of the science 16 (2014) –.

[190] Y. Wang, J. Luo, A. Song, F. Dong, Oats: online aggregation with two-level sharing strategy in cloud, Distributed and Parallel Databases 32 (4) (2014) 467–505, distrib Parallel Databases.

[191] Y. X. Wang, J. Z. Luo, A. B. Song, F. Dong, Partition-based online aggregation with shared sampling in the cloud, Journal of Computer Science and Technology 28 (6) (2013) 989–1011, j Comput Sci Technol.

[192] M. Washisaka, E. Nakamura, T. Takakura, S. Yoshida, S. Tomita, Large-scale distributed data processing platform for analysis of big data, NTT Technical Review 9 (12).

[193] S. A. Wood, A. D. Guerry, J. M. Silver, M. Lacayo, Using social media to quantify nature-based tourism and recreation, Scientific Reports 3, sci. Rep.

[194] H. Wu, A. Yamaguchi, Semantic web technologies for the big data in life sciences 8 (2014) 192–201–.

[195] M. Wu, L. Tan, N. Xiong, A structure fidelity approach for big data collection in wireless sensor networks, Sensors (Switzerland) 15 (1) (2015) 248–273.

[196] D. Xia, Z. Rong, Y. Zhou, Y. Li, Y. Shen, Z. Zhang, A novel parallel algorithm for frequent itemsets mining in massive small files datasets, ICIC Express Letters, Part B: Applications 5 (2) (2014) 459–466, iCIC Express Lett Part B Appl.

[197] D. Xia, B. Wang, Z. Rong, Y. Li, Z. Zhang, Effective methods and strategies for massive small files processing based on hadoop, ICIC Express Letters 8 (7) (2014) 1935–1941, iCIC Express Lett.

[198] J. Xia, C. Yang, Z. Gui, K. Liu, Z. Li, Optimizing an index with spatiotemporal patterns to support geoss clearinghouse, International Journal of Geographical Information Science 28 (7) (2014) 1459–1481, int. J. Geogr. Inf. Sci.

[199] S. Xia, J. Xie, D. Dai, H. Zhang, Q. Nie, S. Kawata, W. Zhang, Kvm combined with hadoop application based-on cpse-bio, Journal of Next Generation Information Technology 4 (3) (2013) 160–166.

[200] H. Xian, K. Madhavan, Anatomy of scholarly collaboration in engineering education: A big-data bibliometric analysis, Journal of Engineering Education 103 (3) (2014) 486–514, j. Eng. Educ.

[201] Q. Xiao, P. Shang, J. Wang, Record-based block distribution (rbbd) and weighted set cover scheduling (wscs) in mapreduce, Journal of Internet Services and Applications 3 (3) (2012) 319–327, j. Internet Serv. Appl.

[202] T. Xiao, Z. Guo, H. Zhou, J. Zhang, X. Zhao, C. Ye, X. Wang, W. Lin, W. Chen, L. Zhou, Cybertron: Pushing the limit on i/o reduction in data-parallel programs, ACM SIGPLAN Notices 49 (10) (2014) 895–908, aCM SIGPLAN Not.

[203] F. Xie, Z. Chen, H. Xu, X. Feng, Q. Hou, Tst: Threshold based similarity transitivity method in collaborative filtering with cloud computing, Tsinghua Science and Technology 18 (3) (2013) 318–327.

[204] G. Xu, W. Yu, Z. Chen, H. Zhang, P. Moulema, X. Fu, C. Lu, A cloud computing based system for cyber security management, International Journal of Parallel, Emergent and Distributed Systems 30 (1) (2014) 29–45, int. J. Parallel Emergent Distrib. Syst.

[205] X. Xu, J. Zhao, G. Xu, Y. Ding, Y. Dong, Dsmc: A novel distributed store-retrieve approach of internet data using mapreduce model and community detection in big data, International Journal of Distributed Sensor Networks 2014, int. J. Distrib. Sens. Netw.

[206] W. Yan, U. Brahmakshatriya, Y. Xue, M. Gilder, B. Wise, p-pic: Parallel power iteration clustering for big data, Journal of Parallel and Distributed Computing 73 (3) (2013) 352–359, j. Parallel Distrib. Comput.

[207] C. Yang, I. Jensen, P. Rosen, A multiscale approach to network event identification using geolocated twitter data, Computing (2013) 1–11.

[208] C. Yang, X. Zhang, C. Zhong, C. Liu, J. Pei, K. Ramamohanarao, J. Chen, A spatiotemporal compression based approach for efficient big data processing on cloud, Journal of Computer and System Sciences 80 (8) (2014) 1563–1583, j. Comput. Syst. Sci.

[209] H. Yang, S. Fong, Improving adaptability of decision tree for mining big data, New Mathematics and Natural Computation 9 (1) (2013) 77–95.

[210] Y. Yang, X. Long, B. Jiang, K-means method for grouping in hybrid mapreduce cluster, Journal of Computers (Finland) 8 (10) (2013) 2648–2655.

[211] S. Yao, J. He, An efficient olap query algorithm based on dimension hierarchical encoding storage and shark, Lecture Notes in Computer Science 8795 (2014) 180–187, lect. Notes Comput. Sci.

[212] Y. Ye, S. Gong, C. Liu, J. Zeng, N. Jia, Y. Zhang, Online belief propagation algorithm for probabilistic latent semantic analysis, Frontiers of Computer Science 7 (4) (2013) 526–535.

[213] X. Yi, F. Liu, J. Liu, H. Jin, Building a network highway for big data: Architecture and challenges, IEEE Network 28 (4) (2014) 5–13, iEEE Network.

[214] J. Yin, J. Zhang, J. Wang, W. C. Feng, Sdaft: A novel scalable data access framework for parallel blast, Parallel Computing 40 (10) (2014) 697–709, parallel Comput.

[215] L. Yong, H. Wenliang, J. Yunliang, Z. Zhiyong, Quick attribute reduct algorithm for neighborhood rough set model, Information Sciences 271 (2014) 65–81, inf Sci.

[216] D. Zhang, T. He, Y. Liu, S. Lin, J. A. Stankovic, A carpooling recommendation system for taxicab services, IEEE Transactions on Emerging Topics in Computing 2 (3) (2013) 254–266, iEEE Trans. Emerg. Top. Comput.

[217] F. Zhang, J. Cao, S. U. Khan, K. Li, K. Hwang, A task-level adaptive mapreduce framework for real-time streaming data in healthcare applications, Future Generation Computer Systems 43-44 (2014) 149–160, future Gener Comput Syst.

[218] F. Zhang, J. Cao, W. Tan, S. U. Khan, K. Li, A. Y. Zomaya, Evolutionary scheduling of dynamic multitasking workloads for big-data analytics in elastic cloud, IEEE Transactions on Emerging Topics in Computing 2 (3) (2013) 338–351, iEEE Trans. Emerg. Top. Comput.

[219] J. Zhang, J. S. Wong, T. Li, Y. Pan, A comparison of parallel large-scale knowledge acquisition using rough set theory on different mapreduce runtime systems, International Journal of Approximate Reasoning 55 (3) (2014) 896–907, int J Approximate Reasoning.

[220] L. Zhang, C. Wu, Z. Li, C. Guo, M. Chen, F. C. M. Lau, Moving big data to the cloud: An online cost-minimizing approach, IEEE Journal on Selected Areas in Communications 31 (12) (2013) 2710–2721, iEEE J Sel Areas Commun.

[221] T. Zhang, L. Cui, M. Xu, A lns-based data placement strategy for data-intensive e-science applications, International Journal of Grid and Utility Computing 5 (4) (2014) 249–262, int. J. Grid Util. Comput.

[222] X. Zhang, C. Liu, S. Nepal, C. Yang, W. Dou, J. Chen, Sac-frapp: A scalable and cost-effective framework for privacy preservation over big data on cloud, Concurrency Computation Practice and Experience.

[223] X. Zhang, L. T. Yang, C. Liu, J. Chen, A scalable two-phase top-down specialization approach for data anonymization using mapreduce on cloud, IEEE Transactions on Parallel and Distributed Systems 25 (2) (2014) 363–373, iEEE Trans Parallel Distrib Syst.

[224] Y. Zhang, M. Chen, S. Mao, L. Hu, V. Leung, Cap: Community activity prediction based on big data analysis, IEEE Network 28 (4) (2014) 52–57, iEEE Network.

[225] Y. Zhang, J. Yang, Optimizing i/o for big array analytics, Proceedings of the VLDB Endowment 5 (8) (2012) 764–775.

[226] Y. Zhao, J. Wu, C. Liu, Dache: A data aware caching for big-data applications using the mapreduce framework, Tsinghua Science and Technology 19 (1) (2014) 39–50, tsinghua Sci. Tech.

[227] Z. Zhao, R. Zhang, J. Cox, D. Duling, W. Sarle, Massively parallel feature selection: An approach based on variance preservation, Machine Learning 92 (1) (2013) 195–220.

[228] E. Zhong, W. Fan, Q. Yang, User behavior learning and transfer in composite social networks, ACM Transactions on Knowledge Discovery from Data 8 (1), aCM Trans. Knowl. Discov. Data.

[229] S. Zhong, D. Chen, Q. Xu, T. Chen, Optimizing the gaussian kernel function with the formulated kernel target alignment criterion for two-class pattern classification, Pattern Recognition 46 (7) (2013) 2045–2054.

[230] Q. Zhou, D. Xiao, Y. Tang, C. Rong, Trusted big data capture and transport architecture for wireless sensor network, Journal of Internet Technology 15 (6) (2014) 1033–1041, j. Internet Technol.

[231] W. W. Zhu, A. Berndsen, E. C. Madsen, M. Tan, I. H. Stairs, A. Brazier, P. Lazarus, R. Lynch, P. Scholz, K. Stovall, S. M. Ransom, S. Banaszak, C. M. Biwer, S. Cohen, L. P. Dartez, J. Flanigan, G. Lunsford, J. G. Martinez, A. Mata, M. Rohr, A. Walker, B. Allen, N. D. R. Bhat, S. Bogdanov, F. Camilo, S. Chatterjee, J. M. Cordes, F. Crawford, J. S. Deneva, G. Desvignes, R. D. Ferdman, P. C. C. Freire, J. W. T. Hessels, F. A. Jenet, D. L. Kaplan, V. M. Kaspi, B. Knispel, K. J. Lee, J. Van Leeuwen, A. G. Lyne, M. A. McLaughlin, X. Siemens, L. G. Spitler, A. Venkataraman, Searching for pulsars using image pattern recognition, Astrophysical Journal 781 (2), astrophys. J.

[232] Z. Zong, R. Fares, B. Romoser, J. Wood, Faststor: Improving the performance of a large scale hybrid storage system via caching and prefetching, Cluster Computing 17 (2) (2014) 593–604, cluster Comput.

[233] H. Zou, Y. Yu, W. Tang, H. W. M. Chen, Flexanalytics: A flexible data analytics framework for big data applications with i/o performance improvement, Big Data Research 1 (2014) 4–13, big. Data Res.