A Systematic Review of Cloud Computing, Big Data and Databases on the Cloud

Completed Research Paper

Alan T Litchfield
Auckland University of Technology
Service and Cloud Computing
Research Lab
alan.litchfield@aut.ac.nz

Jacqui Althouse
Auckland University of Technology
School of Computer and
Mathematical Sciences
jacqui.finlay@gmail.com

Abstract

Cloud computing has emerged as an initiative which offers great promise in improving access to computational resources that would otherwise be unattainable due to sheer cost. While some cloud computing concepts date back to the 1950s, it is recent new cloud architecture and platforms that shape the way that resources are leased using service-based models. However, some confusion exists regarding the relationship between cloud-based models and challenges in managing big data. Some attempt to solve the problem by replacing and upgrading physical infrastructures, while others look to intelligent software to improve the scalability of data analytics. What also remains unclear is the definition and positioning of cloud-orientated paradigms. This is important to establish as it gets to the heart of where the underlying challenges exist in terms of availability, virtualisation, partitioning and distribution, scalability and elasticity, and performance bottlenecks when managing data.

The goal of this systematic review is to provide insight into the current state of cloud computing and big data research. We find that challenges have been gaining momentum in this area from 2008 to 2013. In this study, using a systematic review framework, 129 publications are evaluated. We conclude that the current cloud-computing based frameworks are potentially neglecting fundamental database properties regarding atomicity and durability issues.

Keywords: Cloud Computing, Big Data, Databases, MapReduce, Hadoop, Review, virtualization, distribution, scalability, elasticity, performance

Introduction

A study sponsored by the EMC cooperation shows that the amount of data is doubling every 2 years and estimates that the worlds collective data at 2011 was 1.8 zetabytes (Gantz, & Reinsel, 2011) and we would extrapolate that to 7.2 zetabytes (7.2 trillion gigabytes) by 2015. The new generation of cloud computing solutions is making advances towards leveraging the world's rapidly growing, structured, data supply; however there are existing constraints and limitations in doing so. Researchers are currently facing (and have been facing, over the past three decades) major data management challenges. There is currently a great need for computer scientists to develop cost-effective solutions for data intensive research. Research and development has been relatively slow in this area due to the nature and lack of understandings about database management systems (DBMS), schemas and ontology's (Bell, Hey, & Szalay, 2009). A single perfect data management solution inside cloud computing environments has yet to be designed (Agrawal, Das, & El Abbadi, 2011). Cloud computing as a service has a predicted market worth of $150 (USD) billion in 2014 and $222.5 (USD) billion in 2015 (Carroll, Van der Merwe, & Kotze, 2011).What remains a

challenge is how to make a scalable system that can handle processing large data while efficiently utilising available resources while minimising operational costs.

When data becomes too big to effectively store and analyze, a distributed service, or cloud computing service, becomes a more feasible option. Cloud computing has been introduced as a paradigm of Service Orientated Architecture (SOA) and is derived from a combination of distributed computing, grid computing and virtualisation (through Virtual Machines (VMs)) concepts. This combination of technologies can mean a lot of different things to different people from a range of disciplines and backgrounds. Cloud computing can be the temporary leasing of computer infrastructure, storage and software, over a network, to individuals, businesses and organisations who have requirements for high computational processing power (Foster et al., 2008; Braithwaite and Woodman, 2011). The renting, or leasing, of public computer power is deemed more economically feasible (particularly for Small to Medium sized Enterprises; SMEs) than buying an expensive mainframe (or private cloud) outright (Keung & Kwok, 2012). For an individual there is the benefit of being able to access your data from anywhere, and not having to worry about backups. For researchers cloud computing opens up gateways to data analytics on an enormous scale (Agrawal, Das & Abbadi, 2011). At a taxonomy level, there are three main views of cloud computing, each category represents and defines different layers within the currently available cloud technology stack:

> *Software as a Service (SaaS):* Provides users with access to software applications and services available on the cloud. The physical cloud itself at the software level is relatively hidden. At this level software is deployed and hosted as a service and is accessed over a network (typically the internet) with no requirements for any software installations on the clients-side. The hosting of the software is done by the service provider. SaaS implements a more restrictive model than IaaS as customers are to use only the existing services, defined by the provider, rather than deploying their own.
>
> *Platform as a Service (PaaS):* Provides higher-level programming models and database systems than what is available at the IaaS level. PaaS is also referred to as "Cloudware" and facilitates access of IaaS and SaaS levels through various Application Programmable Interfaces (APIs) in order to develop web applications and services. PaaS, is a less mature layer of cloud computing as it was designed to cater for the challenges faced when managing Quality of Service (QoS) issues by offering developers control over deployed applications and environment configurations without needing to manage underlying infrastructure.
>
> *Infrastructure as a Service (IaaS):* This level provides functionality for hosting raw computing infrastructure (networks, servers, operating systems and storage). An example of a business that operates at IaaS level would be a data centre leasing hardware (storage space) to customers. IaaS as a virtual resource allows a customer to deploy their own software (including operating systems), on top of virtualisation software. At this level a programmer has full access to the virtual machine operating system.

Clouds themselves can be granted access in different ways, depending on their configuration. For example a cloud may be referred to as a "private cloud" if the infrastructure is solely operated and managed by a business or a "community cloud" if computational resources are shared by several organisations. A "public cloud" has its infrastructure available to the general public and is usually owned by a company selling its cloud services. There are also "hybrid clouds" whose infrastructure is made from two or more existing clouds that are bound together by similar infrastructures to increase application portability.

A key theme within these layers is that they all provide services for managing data. We argue that traditional approaches, such as Relational Database Management Systems (RDBMS) are inadequate for handling the ever-growing and complex data challenges as the performance of a RDBMS scales poorly as the amount of data increases (Liu, Xia, Shroff, & Zhang, 2013). Whether or not traditional RDBMS is adequately sufficient for research and industry operations remains to be an open topic (Hacigumus, et al., 2010). The relational model itself is based on passively storing large collections of data with a pre-determined fixed amount of columns. The RDBMS model is designed to cater for intensive input-output (I/O) workloads. However with static constraints an RDBMS may fail to meet the needs of evolving, or dynamic, data. For instance, in the scientific community, operations may require intense computational usage due to complex queries of data that is high in dimensionality and is potentially heterogeneous by

nature. This can cause performance bottlenecks within an RDBMS due to the sheer scale of the data in cloud-based systems (Ward, & Barker, 2013). RDBMSs rely on fully synchronised, static, data elements, however in many stream-orientated applications data arrives at different sections of a cloud asynchronously (Abadi et al, 2003). Since the theme is heavily data-driven the bedrock of this literature is focused on the Databases-as-a-Service (which we consider as merely a sub-concept of SaaS) within the cloud computing environment:

> *Databases a Service (DBaaS):* At the DBaaS level the configuration, scaling, performance, privacy, backup and accessibility issues are managed by the service provider and not the end-users (Mozafari, Curino, & Madden, 2013). There is also a level of awareness that a cloud-DB requires in terms of monitoring querying patterns and data accesses to optimise its use.

While a range of applications have emerged to combat large data analytics problems within the cloud computing context, little research has been done that directly addresses bottlenecks underpinning existing database systems. Studies have acknowledged some of the limitations when dealing with data within a cloud, including: availability, accessibility, data migration, data resilience, scalability, performance, efficient multi-tenancy, privacy and security, yet few have attempted to quantify and investigate these issues and explore potentially inter-related implications. Better understanding of the inter-relations between various cloud components will potentially enable the innovation of new systems based on existing components and systems for enhancing flexibility, extendibility, availability, optimization and cost efficiency.

Research Method

This systematic literature review is based on the procedures and guidelines proposed by Kitchenham (2004) and are used in conjunction with the ideas that constitute an effective literature review by Webster, & Watson (2002). A review protocol is utilized to conduct the search of literature based on the rationale of the research objectives, search strategy, data extraction, literature synthesis and analysis of the findings. Systematic literature reviews offer powerful insights by providing a summary of existing evidence about specific cloud computing issues, identifying gaps in the current state of knowledge and for creating a framework for future research activities.

The contributions of this research are: a description and application of Cloud Computing, Big Data and Databases, and the identification of current research issues and recommendations for future research.

Line of Argument, Hypotheses and Research Questions

Challenges within the cloud-computing domain often appear fragmented as they have been simultaneously addressed in different ways, in different studies, from various perspectives. Kitchenham (2004) describes how Line of Argument Synthesis is applied when, from a set of selected studies, the inference of a whole topic only focuses on part of the issue.

Line of Argument 1: Authors have tended to avoid management issues of large data and therefore have utilised common strategies and tools such as mapReduce and hadoop.

Line of Argument 2: Do current cloud-computing frameworks neglect fundamental ACID database properties?

To achieve the key contributions of this literature review we propose two hypotheses:

Hypothesis 1: Big Data and Cloud Computing challenge classifications may be based on subject or challenge theme while challenge detail also varies.

Hypothesis 2: Big Data and Cloud Computing challenge types may be based on type so that: (i) challenges are addressed from product (software), process (platform) and architectural perspective (infrastructure); (ii) challenges that are addressed from the same perspective are treated as one type; and (iii) challenges are common methods or tools used to accomplish objectives.

We then posit from these two hypotheses, the following research questions:

Question 1: What is the evidence that database architectures are implicated in performance loss when managing very large datasets?

Question 2: What is novel in the evolution of databases in cloud computing?

Question 3: What novel approaches to data modeling in cloud computing exist?

Question 4: What methods and techniques are applied to facilitate data model evaluation?

Question 5: What research challenges exist in the area of database and cloud computing evolution?

Question 6: What is the current ontology for cloud computing systems?

Question 7: What definitions are applied to databases with cloud-based terminology?

Clearly, within the scope of this literature review it is improbable that all these questions are answered and those unaddressed are left for future research articles.

Search Strategy

The Auckland University of Technology (AUT) Search[1] library search system provides a discovery service for searching reliable and credible library content such as digital and print, audio and video, single articles, entire e-journals, as well as a wide range of other formats and sources. The search function delivers a relevancy-ranked list of results. Included in the search results lists are data regarding abstracts, item location and online full text availability.

This systematic review relates to databases, cloud computing and big data challenges. The relation implies that studies are about cloud computing terms and are relevant to the research questions. Using "database", "cloud computing" and "big data" as main key words, a search string is constructed. The search string is constructed using "+" and "-" to indicate inclusive and exclusive phrases, respectively. Out of scope for this work are mobile and security references because the objective is to study how data are managed in the cloud, rather than the application of how data are accessed. A set of inclusion and exclusion criteria are specified based on the scope of this literature and is based on the following search string:

Search String: +"database" and +"cloud computing" and "big data" -"mobile" -"security"

To construct a final dataset and to ensure quality of the review, only peer-reviewed journal articles are accepted. From the article result set, the following details are collected: title, authors, publication year, abstract and the articles full-text. This results in a final collection of primary studies based on the search string used. The main objective of an included study is to present cloud computing and big data challenges, issues and potential future research. Table 1 describes the motivation and rationale of inclusion and exclusion criteria utilised to create a set of articles that is relevant to the current challenges and issues in cloud computing, big data and databases:

---

[1] Previously called Summon, AUT Library Search may be found at:
http://aut.summon.serialssolutions.com/

Participating Publishers: http://www.serialssolutions.com/en/resources/detail/summon-participating-publishers

The comprehensive serials titles list - includes coverage dates and full-text indexing:

http://www.serialssolutions.com/en/resources/detail/summon-serials-titles

Table 1 Inclusion and Exclusion Criteria for Review Protocol

Inclusion Criteria	Exclusion Criteria
A study is mainly about cloud computing, big data and databases issues and contexts *Motivation*: This reviews main objectives are about identifying the current challenges that are related to cloud computing and databases. This means that included articles are relevant to the research questions for this review.	A Study that is not about cloud computing, big data and databases. *Rationale*: Some studies may contain the main keywords within the search term, however they also may be considered irrelevant to the research objectives for this review. This is because these keywords are searched for within an articles text despite the context of the article itself. For example, many papers presented researches on genomic variation and the researchers mentioned that they had used cloud based repositories.
One of the main objectives of a study is to present cloud computing challenges, issues and open questions. *Motivation*: If a paper proposes a challenge, it is expected that a challenge has not been addressed before.	A Study that is about cloud computing in a specific domain. *Rationale*: This review is interested in cloud computing and big data challenges in general, as opposed to specific application domains. This is because challenges in one application domain may be very different from challenges in another application domain.
A study is in the form of a scientific paper. *Motivation*: A scientific paper is commonly viewed as having a higher level of quality and contains reasonable content. The search excludes newspaper articles, magazine articles, newsletters, conference proceedings, and books (as well as e-books).	A study is not in the form of a scientific paper. *Rationale*: A study that is not a scientific paper may not have the guaranteed level of rigour applied to it. In this instance, magazine articles may refer to academic articles but not of themselves rigorously reviewed.

Data Synthesis

The extracted data are synthesized to identify the challenges from the primary study. To provide a summary on the studies that have overlapping themes by translating and comparing each study against a set of similar studies, reciprocal translation is applied. Line of Argument Synthesis is used when we are concerned that what is inferred from a topic as a whole, from a set of selected studies that partially focus on the issue (Kitchenham, 2004). If some of the big data and cloud computing challenges are about a similar theme, then to translate each topic, a reciprocal translation is applied, by examining other topic types. After applying reciprocal translation to all the challenges, then sub-challenges are investigated to determine main themes that are underpinning a set of challenges by using Line of Argument Synthesis.

Key Concepts

Major themes discovered include: availability, virtualization, scalability and elasticity, and performance bottlenecks. Other themes include (amongst others): distribution and partitioning, clustering in cloud environments, technology applications (Hadoop, high performance computing and MapReduce), data management, analytics, comparisons of database architectures, distribution of databases, and database relations.

Availability

Arguments for a cloud-based approach are weak without high levels of availability. Availability as a premise considers the access to IT support, people, IT skills, physical space and time. Availability metrics for determining whether or not to buy a private cloud or rent public cloud time (Keung, & Kwok, 2012) are derived from availability artifacts. Metrics used to predict network traffic during cloud upgrades and maintenance procedures also provide insights into potential cloud failure.

Data are useless and worthless unless they are readily available. However, it is not feasible to have 100% availability of data. The continuation of availability issues leaves some businesses hesitant about moving their internal data to a cloud-based environment. Part of the reason for this is that a lack of availability still presents as one of the major sources of cloud failure. For example, Amazons EC2 infrastructure 2011 failure resulted in a cloud outage in 2011, was caused by a configuration mistake that was due to human error during a network update.[2] In 2009, Google also had a cloud failure during a maintenance phase that involved taking a small number of Gmail servers' offline. In doing so, however, they underestimated the traffic load that would be redirected and the result was that additional request routers had to be setup.[3] Lack of availability of resources can be due to bottlenecks in performance or component failure leading to potentially devastating effects on businesses that rely on cloud-based applications. Availability also introduces challenges to security such as if data, including encrypted data, is leaked to the wrong people or is compromised then it is imperative that appropriate authorities are informed and a contingency plan is set in place.

Another cloud availability challenge is cloud lock-in: a cloud user wants to move their data from one cloud service provider to another but is unable to do so due to complexity and introduced risks when the attempt is planned for or made. This is because of a lack of standards within cloud environment models and resultant interoperability issues subsequently presented. The lack of standards are possibly an expression of the immaturity of the cloud environment and reflects a need for improved understanding of cloud computing and the amount of time it takes to build standards. Additionally, the industry is always going to produce competing and differentiated technologies that create interoperability issues.

Virtualization

At the heart of cloud computing is Virtual Machine (VM) -ware which is a computer system running isolated processes while behaving as a physical system (a multi-processor server can generally run one VM per core). VMs are a common form for allocating and providing computational resources for cloud users. The VM layer resides at the infrastructure (IaaS) level and can be used in conjunction with intelligent software stacks to improve performance and efficiency (Youseff, Butrico, & Da Silva, 2008). Tasks may be processed in parallel on a multi-core VM. Examples of VMware include: EXC/ESXi server, Microsoft Hyper-V R2, KVM (a Linux Kernel-based VM), Xen and Proxmox Virtual Environment. As the field matures, more VMware offerings are being created.

A hypervisor is used to monitor and manage VM operations. There are different VM architectures including hosted architecture (Proxmox and KVM) and bare-metal architecture (ESXi, Microsoft Hyper-V and Xen) (Chang, Tsai, Chen, Lin, & Huang, 2012). The scheduling of multiple VMs on a single CPU core

[2] http://viodi.com/2011/04/22/amazons-ec2-outage-proves-cloud-failure-recovery-is-a-myth/

[3] http://www.theregister.co.uk/2009/02/25/google_gmail_data_centre_fail/

causes issues in overall performance. Xu et al (2012), identifies five main VM allocation strategies for cloud tasks where: 1) the tasks are randomly assigned to one or more VMs (used as a potential benchmark strategy) 2) a group of tasks are assigned to sequentially to one or many VMs, 3) a task is assigned to VM based on the tasks instruction length by descending order, 4) tasks are assigned to a set of VMs that are sorted in ascending order by their execution (CPU) speed and 5) using a greedy stepwise strategy that sorts a set of tasks by their instruction length (in descending order) and sorts a set of VMs by execution speed (in ascending order). Then a time matrix is created based on the execution time for each task on each VM. Tasks are then assigned depending on optimal time frames for processing the task set. While there are various methods to schedule and allocate cloud-based tasks, how to intelligently optimize task schedules using VMs remains an open challenge, due to varying VM speeds and types of tasks (Ward, & Barker, 2013; Xu et al, 2012).

Distribution and Partitioning

In a cloud computing and distributed computing context, a database is often distributed with multiple partitions. Partitioning a database reduces the amount of data read for SQL operations so that overall response time is reduced. A relation within a database can be portioned horizontally (where tables are partitioned by their rows) or vertically (where tables are partitioned by their columns). Horizontal partitions are more commonly used by database vendors as it avoids the need to synchronize distributed tables. Partitioning schemes can also be either static or dynamic. In static partitioning all related rows (or tuples) are stored within a single partition. In dynamic partitioning groupings of tuples are formed based on the efficient processing of workloads, as the size of the data increases (Ahirrao, & Ingle, 2013). Partitioning can also improve availability of resources by ensuring that partitions able to run and respond to some database transactions despite one partition failing (Curino, Jones, Zhang, & Madden, 2010). Distribution and partitioning concepts often fall in line with database ACID properties to ensure that database transactions are processed reliability. These properties consider Atomicity, Consistency, Isolation and Durability. Atomicity refers to transaction failure, where if a transaction fails, the entire transaction is deemed as failing. Transaction consistency is about how a successful transaction will update the state of the database via predefined rules. Isolation refers to how each transaction is treated in isolation to avoid issues using concurrency control methods. Finally, durability is about the ability to retrieve the transaction from permanent storage, once it has been committed to the database.

Other approaches such as graph-databases store records as nodes within a graph and edges between nodes are co-accessed database transactions (Curino, Jones, Zhang, & Madden, 2010). Graph-partitioning algorithms are applied to minimize the number of transactions, or data reads, across a distributed database. While there are benefits to partitioning there are also constraints in assuring consistency, availability and partition tolerance. This is known as the CAP theorem, or Brewer's theorem, which states that it is impossible to: guarantee 1) consistency, where data arrives to all nodes within a network at the same time; 2) availability, that every request has a failed or successful response; and 3) partition tolerance, where a system will be able to operate despite any component failures or lost requests. Ahirrao, & Ingle (2013) report that while researchers have developed some novel approaches to partitioning to improve scalability, there is no technique available that will effectively partition the database based on access patterns.

Scalability and Elasticity

A DBMS within a cloud environment is required to efficiently manage different databases (tenants), schemas, workloads, data accesses and resources. Some of these tenants can start relatively small while others grow at unpredictable rates. Managing multi-tenancy to ensure a balance of good performance at a low cost requires efficient use of resources at various peak times for different tenants. Traditional RDBMSs lack the ability to efficiently adapt transaction processing at ever-changing peak-times for multiple tenants (Das, Agrawal, & El Abbadi, 2013). Large scale database systems attempt to overcome the limitations, due to lack of scalability of traditional RDBMSs. Examples of large (distributed) scale DBMSs include: Google's BigTable (2004), PNUTs (2008), Amazon DynamoDB (2012), G-Store (2010), Megastore (2011), Deuteronomy (2011), Relational Cloud (2011), Cloud TPS (2011), Cloud SQL Server (2011), ElasTraS (2009) and Albatross (2011).

Scalability within the cloud-computing paradigm refers to the static property of cloud and how the system will perform based on a static configuration. Scalability itself is comprised of two primary notions: Firstly, currently a virtually unlimited level of scalability is deployed with the key-value stores a DBMS utilizes to provide higher-levels of functionality when providing transactional access to multiple database entities (Agrawal, Das, & Abbadi, 2011). The second notion is to leverage the DBMS architecture itself used in conjunction with key-value stores. Scalability refers to the extent that databases within the cloud environment are able to elegantly handle growing amounts of work with the addition of resources (hardware) (2011). Hardware resources can scale the system either vertically (scaling-up) or horizontally (scaling-out). By adding resources to a single node (computer) and where processor power and memory capacity are increased to improve the performance of virtualization, scaling-up is achieved. Adding more nodes to a distributed system facilitates scaling out.

Elasticity is essential for cloud systems that are leasing their resources on a pay-per-use basis. Elasticity refers to the ability to minimize operational cost while ensuring optimal performance regardless of computational workloads (Das, Agrawal, & Abbadi, 2013). During low workload periods, a cloud-based system should consume minimal resources to minimize operational cost (for example, by improving energy efficiency). However, powering down servers introduces a potential threat to service availability, and bringing servers back online can be relatively expensive if power-resources are predicted incorrectly. Curino, Jones, Zhang, & Madden (2010) describe how elasticity is achieved through live database migration, that is where data from one server are moved to another before shutting the previous server down. For this to be effective, the migration process must have low impact on overall service performance with minimal interruption. Shared disk and shared nothing are the two most common cloud database architectures used in data migration. Shared disk architectures are used for creating abstractions and replications of data (Bigtable, HBase, ElasTraS use shared disk architecture). Shared nothing architecture stores a persistent database image for multiple tenants. This requires all database components to be migrated between nodes.

We have found few studies that detail a comprehensive comparative analysis of the performance of such storage systems. This is again partially because the systems themselves are immature. Systems like Bigtable, PNUTS and Dynamo tend to support single-row transactions and key-value look up functions rather than managing the scale-out transactions that occur due to the nature of multitenancy. Other solutions (for example, Megastore and ElasTras) attempt to efficiently cater for a large number of tenants. Managing unpredictable resource sharing is an open topic. It is challenging to find empirical studies that have a detailed analysis of potential performance bottlenecks caused by current scalability and elasticity frameworks.

Performance Bottlenecks

Bottlenecks are generally always going to be present, in one form or another within a computer system as there is usually some component(s) that restricts the overall performance. In cloud computing, virtualization introduces a bottleneck due to the randomization of I/O operations (CPU Queuing). Other bottlenecks occur due to lack of server power, I/O timeouts due to server overhead, database locking, lack of storage capacity and cache flushing. These bottlenecks can cause reliability issues and there does not appear to be a "silver bullet" that will improve service reliability. There are many causes for potential failure. Typically, one solution is to replace older (slower) infrastructure with newer (faster) infrastructure. Other solutions require installations of intelligent software to optimize workloads within the VM ware (for example, Hadoop using the MapReduce programming model as a "big data" processing tool). Bottlenecks such as connectivity timeouts, workload surges, input errors and client-side component failures also exist at the user-end, which are outside of the cloud services control.

One of the challenges in cloud computing, in terms of data processing, is the movement of data from one place to another. Typically within a cloud, you will have n-Servers, storing n-Partitioned rows of data on n-Databases. The physical location of the data (where they are stored) remains relatively unknown until it needs to be located. Data arrive at stochastic intervals from n-networks and n-servers within the cloud, but where updates are performed is unknown. Thus, servers are required to communicate with each other in order to find where the update is to occur. The total amount of time taken for the servers to communicate and perform an update can be computationally expensive. This implies constraints exist

within the network because of the large volume of data moving from machine-to-machine within a finite bandwidth. Generally, latency has a role in network efficiency and performance loss. Despite the use of sophisticated compression algorithms to minimize the volume of data, volume and frequency remains an open issue.

Bottlenecks also reside within database architectures at various levels. For example, in RDBMS tables, rows and columns are fixed and this becomes a limitation when dealing with data that are more varied than that of traditional information systems. Mozafari, Curino, & Madden (2013) argue that there is a need for models and tools that effectively and efficiently predict resource allocations before databases become suitable for operating in a cloud environment.

Findings

Presented in Table 2 is the number of journal articles, magazine articles, conference proceedings and books (including eBooks) published over the years between (and inclusive of) 2009 and 2013 from the utilised search strategy. Regardless of the source type, there appears to be a steady increase in studies of databases, cloud computing and big data.

Table 2 Summary of Studies According to Content Source Type

	2009	2010	2011	2012	2013	Total
Journal Articles	3	7	13	61	45	**129**
Magazine Articles	2	3	11	44	23	**83**
Conference Proceedings	0	5	8	27	23	**63**
Books / eBooks	0	0	0	2	2	**4**
Total	**5**	**15**	**32**	**134**	**93**	**279**

Overview of the Journal Articles

In total there were 129 search results from the adopted search strategy, which is a substantial decrease from simply using the search term: "cloud computing" alone (refer to search results shown in Figure 1 comprising of 20,440 journal articles,). 119 full-text articles were retrieved from the found set. Using the inclusion and exclusion criteria from the literature review protocol, by removing obscurities the results were filtered. The result was 99 articles considered relevant for this review. Figure 2 shows the distribution of journal articles published per year (as at 2013) along with the trend-line. Surprisingly, while the key concepts of cloud computing have historical precedents that date back to the 1950s there were no publications found prior to 2009. The upwards slope of the trend line implies that there is an increasing amount of research on cloud computing, data and databases. There were 61 journal articles published in 2012, a substantial increase from the 13 articles published in 2011.

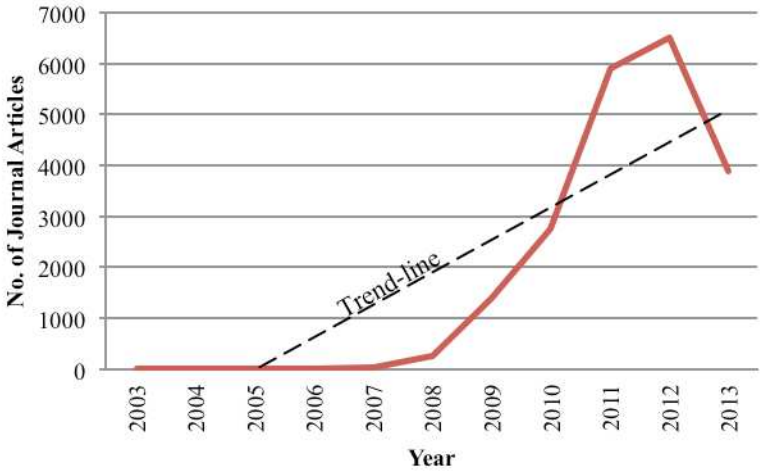


Figure 1. Distribution of Journal Articles Published per Year about "Cloud Computing"

With reference to the number of articles published over the same period, the obvious difference between Figure 1 and Figure 2 implies a significantly smaller amount of research about the connection between cloud computing, big data and database issues. However, the more rapid increase of the number of papers related specifically to databases and big data challenges within the cloud computing context suggests the recent interest is growing.

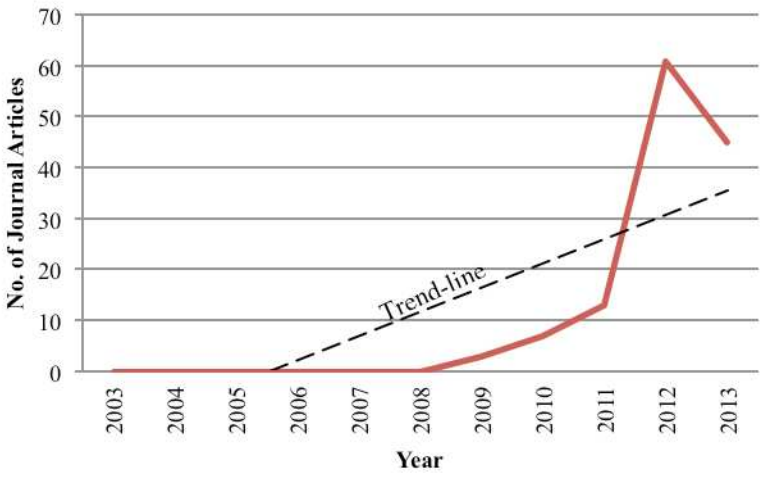


Figure 2. Distribution of Journal Articles Published per Year Utilizing the Proposed Search Strategy

When changes are continuous and complexity increases then to reduce complexity, there are more maintenance efforts. With "big data" and "cloud computing" the higher the expectations, the more work is dedicated to identify challenges. Consequently, with more insights about challenges then more studies are likely to be dedicated to overcoming existing limitations. The increasing number of cloud computing challenges appearing in journals indicates incremental insights into the field and shows that challenges are attracting attention within the research community. In Figure 3 are additional "common search terms" and their number of occurrences within the search results. These search terms are generated from the search strategy as additional search meta-data. Here it is noted that applications of cloud computing

occur in multiple disciplines and primarily ranging from areas in advertising, marketing, bioinformatics, genomics and biology.

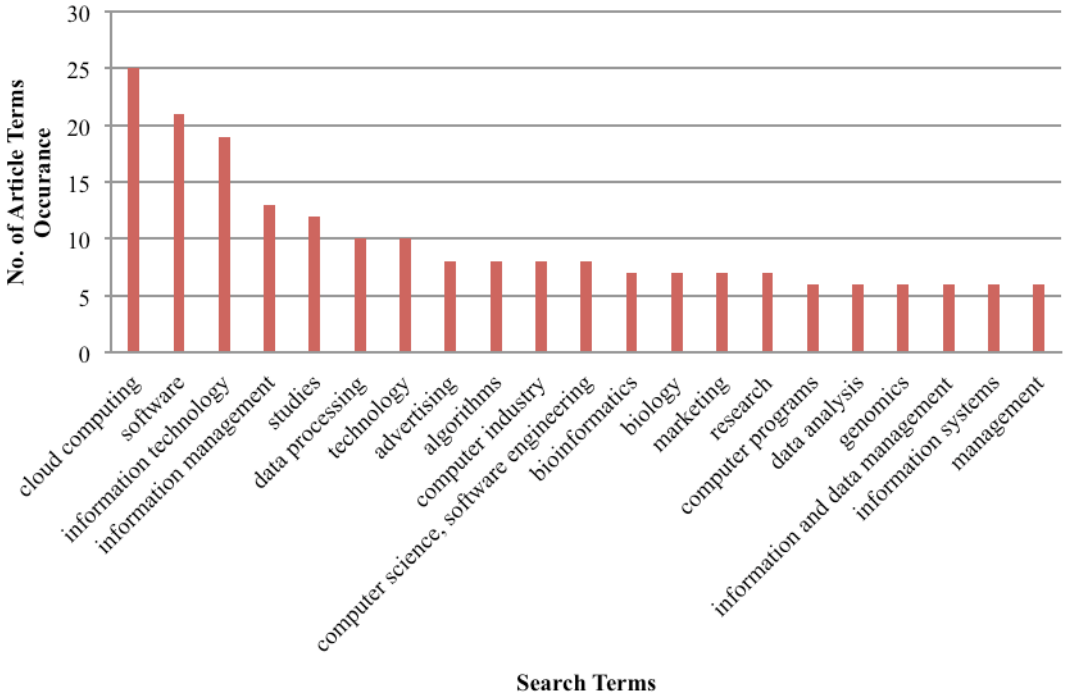


Figure 3. Number of Journal Articles by Search Terms

Degree of Article Coverage

To identify potential themes and sub-themes existing within the found literature, a word frequency query is executed. Frequently used words that are not deemed relevant to the context of this review are removed. By utilising the inclusion and exclusion criteria for selecting articles, the selection is made. The same set of criteria is applicable to searching for themes within the literature as the focus appropriately shifts towards a finer level of granularity. Table 3 details the number of sources (articles) and number of references made within those sources (the number of times a keyword is used) for the three major themes identified. While the literature themes and sub-themes are singularly presented, it is noteworthy that in actuality they are very much interrelated.

Table 3 Number of Sources and References by Search Terms

Theme	No. of Sources	No. of References
Data	115	6310
Cloud	99	1386
Databases	75	252

Data as a theme has the highest number of sources, followed by cloud and databases. This suggests that the key issues in cloud computing involve the management of data and we would surmise that this is equivalent to big data given the range of articles surveyed. The data theme contains three primary sub-

themes, illustrated in Figure 4, that relate to: 1) how data are aggregated, moved and merged, 2) methodologies for big data analytics and 3) how meaning is extracted from data (semantics). Table 4 presents the number of references made within article sources about data related themes. The sub-themes are identified via searching for synonyms of "aggregate", "Analytics" and "Semantic" because these are terms that frequently occur within the sample. Eliminated from these themes were discussions about the analysis of specific application domains such as, if analytics was contextually based on the analysis of DNA trends, as opposed to being focused on cloud-based frameworks analytics (then the reference is not counted). This provides evidence that currently suggests that analysis of big data presents a major challenge within the cloud computing research community. What is interesting is that the issues underpinning analytical methods for big data appears to relate to how data are aggregated and interpreted.

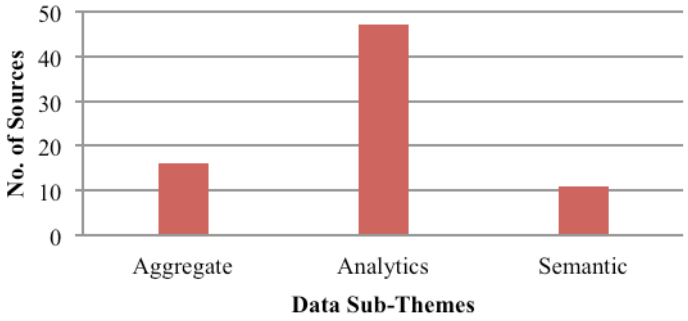


Figure 4. Distribution of Articles Sources for Sub-Themes of Data

Table 4 Number of References Made within Article Sources for Data Themes

Theme	No. of Source References
Data	**2522**
Aggregate	54
Analytics	107
Semantic	41

The second major theme refers to the cloud component itself. There are several reoccurring sub-themes surrounding this topic, shown in Figure 5. This includes: 1) issues relating to bandwidth (including data transfers and impedance at an infrastructure, IaaS, level), 2) clusters, or cluster technologies and the distributed database systems that reside within them, 3) platform (PaaS) related issues, 4) performance challenges, 5) Service Orientated Architecture (SOA) challenges, 6) technologies including cloud computing frameworks like Microsoft's Dryad or the open source Apache project, Hadoop. In addition to this, High Performance Computing (HPC), the Internet, MapReduce, Storage Area Networks (SAN) and Storage-Defined Networks (SDN) were also commonly referenced technologies. Subsequently, shown in Table 5 is the number of references to cloud computing related themes within the articles sourced.

Commonly occurring themes within performance challenges relate to issues of lack of availability and performance bottlenecks. Both these issues concern data constraints by various technologies that exist at infrastructure, platform and software levels. The adoption of Hadoop (using MapReduce as a programming model) as a framework is frequently referred to, however SAN and SDN frameworks are rarely discussed.

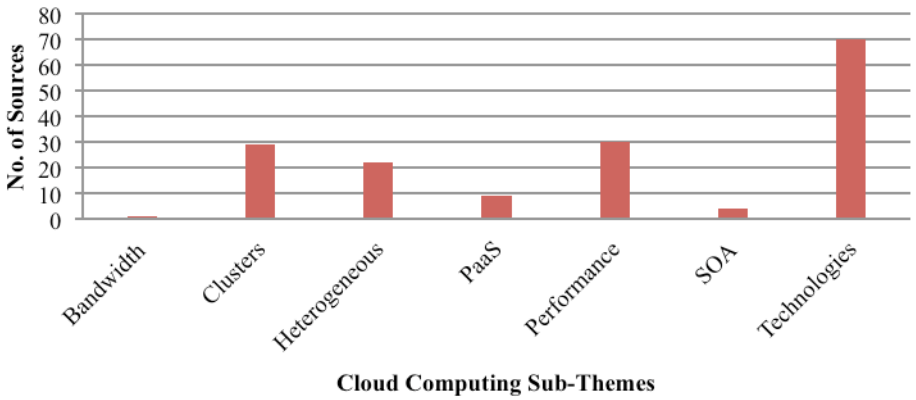


Figure 5. Distribution of Articles Sources for Sub-Themes for Cloud Computing

Table 5 Number of References Made within Article Sources for Cloud Computing Themes

Theme	No. of Source References
Clouds	1386
Bandwidth	1
Clusters	115
Heterogeneous	78
PaaS	17
Performance	59
• Availability	29
• Bottleneck	26
SOA	13
Technologies	952
• Dryad	48
• Hadoop	226
• HPC	354
• internet	76
• MapReduce	243
• SAN	4
• SDN	1

The third major theme is databases, which also consists of several sub-themes (Figure 6). Commonly occurring database themes include: 1) architecture, where only the references to architectures that relate to database technologies are considered, 2) distributed database systems, 3) elasticity research issues

about database design and architecture (not including references to Elastic Compute Cloud, EC2), 4) partitioning challenges, 5) database challenges related to ACID properties, 6) relationships within relational cloud databases, and 7) research challenges relating to the scalability of cloud-based technologies. Table 6 presents the number of references made within article sources about database related themes.

Surprisingly, there are only a few references to database elasticity and ACID properties, which we regard as significant in both cloud computing and database technologies. Overall, there were six references to the keyword "properties" within the database context. There were a small number of references to Completion (4 article sources) and Isolation (2 article sources). There is no mention (zero references) to Atomicity and Durability. There is also no mention of key database themes that focus on partitioning, an inherently important factor when configuring cloud databases. This implies a gap in the literature. It is possible that research is not adequately challenging fundamental database theories within the cloud computing context.

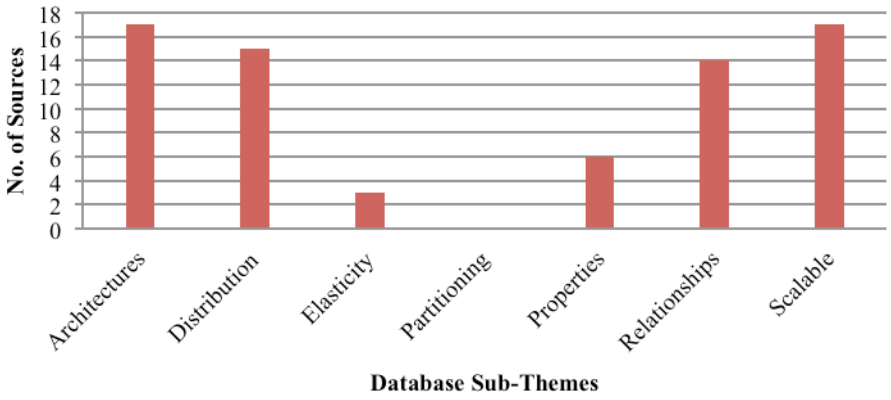


Figure 6. Distribution of Articles Sources for Sub-Themes for Databases

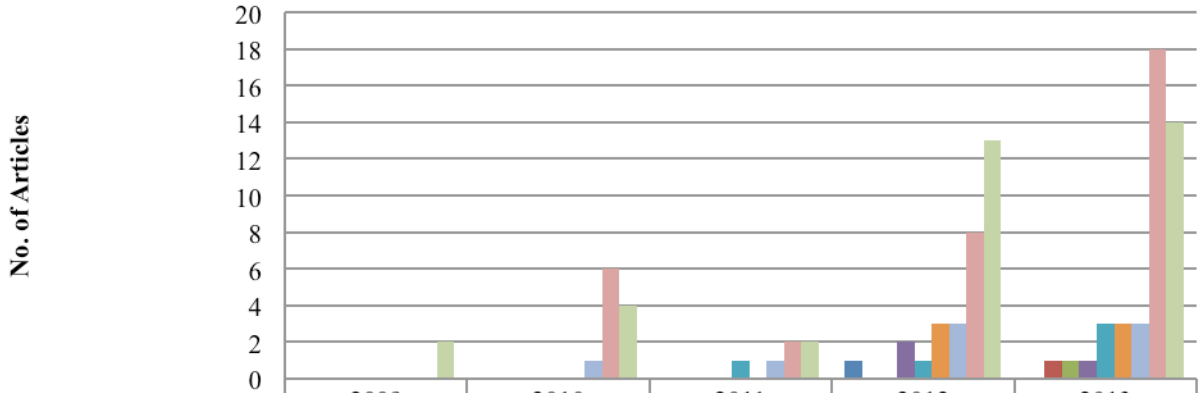
Table 6 Number of References Made within Article Sources for Database Themes

Theme	No. of Source References
Databases	243
Architectures	37
Distribution	21
Elasticity	5
Partitioning	0
Relationships	34
Scalable	24
Properties:	8
• Atomicity	0
• Completion	6
• Isolation	2
• Durability	0

An analysis groups themes together on the basis that they have words in common using the Pearson correlation coefficient. It is observed that "data" as a central theme is inter-related to challenges focusing on bottlenecks, heterogeneity, scalability, analytics, semantics, relationships, entities, data aggregation and distributed databases. The outer-themes represent challenges that are not addressed well in the literature. These issues include databases, clusters, partitioning and SDN challenges. In addition to this, corresponding sub-themes also appear to be outlier nodes. These include atomicity and durability from database ACID properties as well as issues surrounding referential database relationships and elasticity. Intuitively, issues surrounding elasticity, atomicity and durability appear within close proximity to each other. There is some confusion as to why databases, atomicity, elasticity and durability remain removed from the data theme. In addition to this, the database theme is also distinct from its own ACID properties. However, this could be partially due to the decreased number of sources that reference these properties.

Tools Methods and Techniques

To gain insights into tools and methods utilized currently, we conduct a text-based search. Figure 8 shows the number of references made to various cloud computing resources. One of the limitations of this search is that the "No. of Articles" field may, in some cases, exist as a reference within an article and not actually be a tool utilized within an article's study. However, the search does show that cloud technologies such as Amazons EC2 and Hadoop are popular cloud computing frameworks and that this is recent. For other cloud computing resources such as Enomaly, Joynet and Zoho there were zero reference occurrences. Overall, the most popular frameworks included Amazon EC2 (with a total of 35 articles) and Hadoop (34 articles), followed by Microsoft's Dryad (8 articles) from the search criteria used.



	2009	2010	2011	2012	2013
Eucalyptus	0	0	0	1	0
Puppet	0	0	0	0	1
Globus Nimbus	0	0	0	0	1
OpenNebula	0	0	0	2	1
Cloudera (Hadoop)	0	0	1	1	3
Windows Azure	0	0	0	3	3
Microsoft Dryad	0	1	1	3	3
Hadoop	0	6	2	8	18
Amazon EC2	2	4	2	13	14

Figure 8. Cloud Computing Resource References

Discussion

The results of the literature survey suggest an unveiling of innovation that is gaining momentum. This trend is not uncommon within the IT industry where "gale[s] of creative destruction" (a term coined by economist Joseph Schumpeter in the 1940s) increase economic growth by rearranging resources to create more value while transforming global markets and cultures through entrepreneurship (Fremdt, Beck, & Weber, 2013). The rearrangement of computing resources have enabled IT-inexperienced businesses, relying on ad-hoc methods for managing IT infrastructure, to offload this responsibility to cloud service providers (Braithwaite, & Woodman, 2011). There are various views on what cloud computing actually means and while there are defined layers to understanding cloud computing concepts, these layers often overlap and it can create a degree of confusion within the literature.

There appears to be a lack of clarity and precision in the use of cloud computing terms. The lack of clarity reduces the degree of interpretability of cloud-based studies and makes communication of cloud terminology difficult. As a result, a challenge in determining the comparability of the results is introduced and potential learning opportunities is lowered. For example, scalability and elasticity are often used interchangeably when in actuality they represent two separate concepts. The lack of precise terminology could be partially due to creative destruction patterns observed within current literature trends. While there is much competition towards driving the world's leading state-of-the-art cloud computing frameworks, there are gaps in our understanding of the fundamentals that underpin them. It is therefore recommended that any terminology used is clearly defined, and mixing terminologies in future studies ought to be avoided.

We have observed that in the application of adaptive systems, from the very large and rapidly growing amount of rich, unstructured data and on-demand computational resources, issues are being introduced. While there appears to be a focus on large-scale data analytics, there are conflicts between stability, accuracy, robustness, trust and control. Providing these elements while also offering elasticity over a wide set of conditions requires a system to adapt over time (Zliobaite et al., 2012). Thus, there is a degree of transparency required for adaptive mechanisms and research efforts towards automating data pre-processing, prediction and feedback within a cloud computing context is encouraged. Currently, there appears to be a lack of experimental data on measures of success and failure of cloud computing frameworks and how they perform under a combination of various database schemas (and systems), configurations and VMs. While bottlenecks are acknowledged within the literature, it appears that few studies actually perform an in-depth comparative analysis to determine where they exist and their implications and relationships to big data Atomicity, Consistency, Isolation, Durability and performance.

Within the literature, the data-theme no longer exists in isolation and has been combined in various ways to extract knowledge. However, when data are transformed by abstraction and aggregation (via programming models such as MapReduce) to create smaller subsets for analytical purposes, the bigger picture is potentially lost. Data filtering mechanisms may classify data as "noise" and potential anomalies are ignored, where in reality these values have the potential to provide new and interesting insights into existing patterns. In addition to this, these data filtering methods often lack schema support for other database systems functions that could enhance performance.

Threats to Validity

For this systematic literature review the selection of primary studies and data extraction processes contain subjective measures that may introduce a degree of bias within the results of the findings. The three major areas of validity threats include: construct validity, internal validity and external validity.

Construct validity refers to the extent to which inferences may be correctly construed from the findings. Since the review was undertaken by its designers, there are very few threats to construct validity. Thus, the chance that theoretical concepts have been misinterpreted is reduced. The objective of this review is to explore current challenges faced within cloud computing and databases. A systematic review methodology has been adopted so that cloud computing and big data challenges are defined as current research issues. As a result, the reader of this article can arrive at the same interpretations of this study's findings.

Therefore, the study may be replicated. Based on this review's protocol, other researchers can perform an up-to-date replication of this study in the future.

Internal validity is concerned with the extent to which bias and errors are minimized. To reduce internal bias the review protocol requires an explicit definition of the research questions, search strategy, study selection with precise inclusion and exclusion criteria, as well as data synthesis methods. To ensure the study's rigor, a secondary researcher has conducted quality assurance by also evaluating the data set. While this minimizes the potential subjective influence of the results, it does not eliminate it.

One threat to the design of the review involves two search terms ("mobile" and "security") that are applied. In an attempt to reduce irrelevant studies from inclusion, some occurrences may have been included. For example, it is possible that an article may propose new big data and cloud computing challenges but as a research objective, it may be ambiguous.

External validity is about the generalizability of the results of the study. The scope of the review is restricted to academic scientific articles searched by AUT Library's electronic search engine, Summon. Since other forms than scientific articles are not considered, then the completeness of the results is threatened. We assume that big data challenges in cloud computing frameworks found within the academic domain also occur in cloud computing practice, regardless of who uses them.

Conclusion

The main contributions of this review include an overview and classification of the major data and cloud computing challenges being recognised within the research community. This article presents the results of the systematic review as well as the research methodology followed. The results imply that challenges to cloud computing based frameworks exist because of a lack of application of fundamental database properties. More specifically database ACID properties surrounding Atomicity and Durability issues.

There is a need for research that explores the performance limitations of data management and RDBM theory within a cloud-computing context. In doing so the relationships between the current challenges need to be further examined and their detrimental impact on data management in terms of I/O and Extract-Transform-Load (ETL) related tasks evaluated. What constitutes effective cloud-based performance and quality criteria provides an area that requires further investigation. While there is much discussion about availability of resources via service agreements, there appears to be less emphasis on dynamically occurring bottleneck issues.

Cloud computing challenges have various meanings and that have been applied in different studies. When these meanings are clustered, the fundamental challenges consistently focus on the data theme. Regardless of the application domain, inter-dependencies between big data challenges arise, for example there are relationships between specific research domains and the collection of large data sets, the engineering requirements to process and store large volumes of data, and performance and stability issues when managing large numbers of transactions. While the quantity of literature surrounding cloud computing, databases and big data research is gaining momentum, it appears that the practical operation of cloud-based systems is the main focus of the research as opposed to building new theories addressing big data challenges. We would encourage researchers to consider that aspect in particular. The field of cloud-based research is young and before the field is overrun by derivative explorations of past technologies that have been repurposed or redeveloped for new applications, we would strongly support studies that challenge those approaches with a view towards building technological advances for the future.

References

Agrawal, D., Das, S., & El Abbadi, A. 2011. "Big data and cloud computing: current state and future opportunities". In *Proceedings of the 14th International Conference on Extending Database Technology* pp. 530-533.

Agrawal, D., El Abbadi, A., Das, S., & Elmore, A. J. 2011. "Database scalability, elasticity, and autonomy in the cloud". In *Database Systems for Advanced Applications*. Springer, Berlin Heidelberg. pp. 2-15.

Ahirrao, S., & Ingle, R. 2013. "Scalable transactions in Cloud Data Stores". In *Advance Computing Conference (IACC), 2013 IEEE 3rd International*. pp. 116-119.

Bell, G., Hey, T., & Szalay, A. 2009. "Beyond the data deluge". *Science*, (323:5919), pp1297-1298.

Braithwaite, F., & Woodman, M. 2011. "Success Dimensions in Selecting Cloud Software Services". In *2011 37th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA),* pp. 146-154.

Carroll, M., Van der Merwe, A., & Kotze, P. 2011. "Secure cloud computing: Benefits, risks and controls". In *Information Security South Africa (ISSA)*, pp. 1-9.

Chang, B. R., Tsai, H. F., Chen, C. M., Lin, Z. Y., & Huang, C. F. 2012. "Assessment of Hypervisor and Shared Storage for Cloud Computing Server". In *2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications (IBICA)*, pp. 67-72.

Curino, C., Jones, E., Zhang, Y., & Madden, S. 2010. "Schism: a workload-driven approach to database replication and partitioning". *Proceedings of the VLDB Endowment*, (3:1-2), pp 48-57.

Das, S., Agrawal, D., & El Abbadi, A. 2013. "ElasTraS: An elastic, scalable, and self-managing transactional database for the cloud". *ACM Transactions on Database Systems (TODS)*, (38:1), p 5.

Foster I, Yong Zhao, Raicu I, Lu S. 2008. "Cloud computing and grid computing 360-degree compared". In *Grid computing environments workshop, 2008. GCE '08grid computing environments workshop*, Austin, TX, pp 1–10

Fremdt, S., Beck, R., & Weber, S. 2013. "Does Cloud Computing Matter? An Analysis of the Cloud Model Software-as-a-Service and Its Impact on Operational Agility". In *2013 46th Hawaii International Conference on System Sciences (HICSS)*, pp. 1025-1034.

Gantz, J., & Reinsel, D. 2011. "Extracting value from chaos". *IDC iView*, pp 1-12.

Hacigumus, x, mu, s, H., Tatemura, J., Wang-Pin, H., Jafarpour, H. 2010. "CloudDB: One Size Fits All Revived". In *2010 6th World Congress on Services (SERVICES-1)*. doi: 10.1109/SERVICES2010.96

Hevner, A. R., March, S. T., Park, J., & Ram, S. 2004. "Design science in information systems research". *MIS quarterly*, (28:1), pp 75-105.

Keung, J., & Kwok, F. 2012. "Cloud Deployment Model Selection Assessment for SMEs: Renting or Buying a Cloud". In *Proceedings of the 2012 IEEE/ACM Fifth International Conference on Utility and Cloud Computing,* pp. 21-28.

King, R., McLeod, D. 1985. "A Database Design Methodology and Tool for Information Systems". *ACM Transactions on Information Systems*, (3:1), pp 2 - 21. doi 10.1145/3864.3869

Kitchenham, B. 2004. *Procedures for performing systematic reviews*. Keele, UK, Keele University.

Liu, J., Xia, C., Shroff, N., & Zhang, X. 2013. "On distributed computation rate optimization for deploying cloud computing programming frameworks". *ACM SIGMETRICS Performance Evaluation Review*, (40:4), pp 63-72. doi: 10.1145/2479942.2479950

Malecha, G., Morrisett, G., Shinnar, A., & Wisnesky, R. 2010. "Toward a verified relational database management system". In *ACM Sigplan Notices* (45:1), pp. 237-248.

Moschakis, I. A., & Karatza, H. D. 2012. "Evaluation of gang scheduling performance and cost in a cloud computing system". *The Journal of Supercomputing*, (59:2), pp 975-992.

Mozafari, B., Curino, C., & Madden, S. 2013. "DBSeer: Resource and Performance Prediction for Building a Next Generation Database Cloud". In *CIDR*.

Roussopoulos, N., & Yeh, R. T. 1984. "An adaptable methodology for database design". *Computer*, (17:5), pp 64-80. doi: 10.1109/MC.1984.1659139

Stonebraker, M., Abadi, D., DeWitt, D. J., Madden, S., Paulson, E., Pavlo, A., & Rasin, A. 2010. "MapReduce and parallel DBMSs: friends or foes?". *Communications of the ACM*, (53:1), pp 64-71.

Stonebraker, M., Ailamaki, A., Kepner, J., & Szalay, A. 2012. "The Future of Scientific Data Bases". *2012 IEEE 28th International Conference on Data Engineering*.

Suciu, D. 2013. "Big Data Begets Big Database Theory". In *Big Data* (pp. 1-5). Springer Berlin Heidelberg.

Ward, J. S., & Barker, A. 2013. "A Cloud Computing Survey: Developments and Future Trends in Infrastructure as a Service Computing". *arXiv preprint* arXiv:1306.1394.

Webster, J., & Watson, R. T. 2002. "Analyzing the Past to Prepare for the Future: Writing a Literature Review". *MIS quarterly*, (26:2).

Xu, X., Hu, H., Hu, N., & Ying, W. 2012. "Cloud Task and Virtual Machine Allocation Strategy in Cloud Computing Environment". In *Network Computing and Information Security*. Springer Berlin Heidelberg pp. 113-120.

Youseff, L., Butrico, M., & Da Silva, D. 2008. "Toward a unified ontology of cloud computing". In *Grid Computing Environments Workshop, 2008. GCE'08*. pp. 1-10. IEEE.

Zliobaite, I., Bifet, A., Gaber, M., Gabrys, B., Gama, J., Minku, L., & Musial, K. 2012. "Next challenges for adaptive learning systems". *SIGKDD Explor. Newsl.*, (14:1). doi: 10.1145/2408736.2408746. pp 48-55.