View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CiteSeerX

Information quality in Cloud Computing: improvement or deterioration.

Javier Flores

University of Texas Pan American,
1201 W. University Drive, Edinburg, TX, 78539.
Phone (956)467-6293 jflores6@broncs.utpa.edu

Abstract

The convergence of previous technologies such as Virtualization, Grid computing, Distributed computing, Web Services and Service Oriented Architecture, have come with a good timing contributing to build up Cloud Computing. The services usually found in local data centers are now offered to be delivered from outside of the company, with the possibility of a virtually unlimited quantity of computer resources. But we need to consider how beneficial or unfavorable this new setting is, for issues not completely solved in the past, such as Data Quality. The objective of this study is to form a simplified model that can be used to identify the impact of cloud computing on information quality attribute, being able to measure it, and with this look for improvements in the system quality overall. The results can be useful to people involved in information technology management, data management personnel, system development, project management, and data architecture.

Keywords: Information quality, data quality, cloud computing, relative weighted method.

Introduction

An interest among many fields in science and business has been growing up these past years around the confluence of varied technologies that form a cloud computing environment. Many academic people have explored for the definition of a common platform, to the possible implications to their respective working arena. How this new conception can help their research, by providing a much bigger platform where they can develop their work, together with a more flexible and faster provisioning system, and lower investment, to set the correct environment.
Together with this, experts are working around data quality issues in computer information systems, arriving to some answers and solutions during a period of time of more than 20 years, but they establish the need for additional efforts on this field (Madnick and Wang, 2009).
As mentioned, efforts towards finding a definition for cloud computing has been looking back to not only previous work on cloud computing (Vaquero, et al., 2009), but also on the relations of previous technologies that are key components of this new concept (Xu, K., et al., 2009), mainly grid computing, virtualization, distributed computing, service oriented architecture, web services (Zhang and Zhou, 2009).

Even though cloud computing is a term started to be repeated in dialogues and writings amongst academicians and business people, the technology that makes it possible have been studied for a longer period of time. There are works around the similarities and differences of this new paradigm and its previous technologies, in order to speed up prototypes to production systems (Foster, I., et al., 2008; Korri, T., 2009). And sometimes studies even mention confusion around the cloud computing technology itself, but mainly with its interrelationships among their components (Foster and Tuecke, 2005, Xu, M., et al. 2009). Attempts have been made to come with akin terms for concepts around cloud computing, its components from a product and service point of view (Klems, et al., 2008; Creeger, 2009). And one definition in a study that consider old and new concepts, such as virtualization, distributed and grid computing, establish that "cloud computing can be considered a new computing paradigm that allows users to temporary utilize computing infrastructure over the network, supplied as a service by the cloud-provider at possibly one or more levels of abstraction" (Youseff, et al., 2008).

Regarding data quality, studies to assess data reliability in the area of information systems (Agmon and Ahituv, 1987) were performed as first steps towards the understanding of the data representation of the reality. Then, definitions around the concept have been studied for more than 20 years (Redman, 1995). And work to establish the data quality dimensions (Wand and Wang, 1996) and data quality categories (Strong, et al., 1997) were foundations for many future studies (Madnick, et al, 2009). Subsequent works, mainly in the Massachusetts Institute of Technology, provided a view of data: "…enterprises and managers must understand the properties of data and manage them as resources…data, unlike many resources, are intangible, easy to copy and transport, and renewable." (Levitin and Redman, 1998). And "A key insight…data is, in fact, a product (or by-product) manufactured my most organizations, it was not treated nor studied as such" (Madnick and Wang, 2009).

Since cloud computing is such a new paradigm, academic people have started to explore lately this concept, and a lot more researchers have worked out the importance of data quality. Some of these studies, and their contributions include:

- Agmon and Ahituv (1987) and their development of three measurements for data reliability, originally from the field of quality control, and applied to Information Systems. The measures are: internal reliability (commonly accepted), relative reliability (compliance to user requirements), and absolute reliability (level of similarity of data to reality)
- Redman (1995) provided an applicable view of data as competitive advantage in business, giving an insight to data quality issues, and strategies to be implemented to improve data and information quality.
- Wand and Wang (1996) worked, despite the lack of agreement on a set of data quality dimensions, and even data quality definitions, to establish four data quality dimensions. These dimensions reference data as: complete, unambiguous, meaningful, and correct. The analysis considered the view obtained by direct observation and the view obtained from an information system.
- Strong, et al. (1997) and their development to set data quality definitions, considering data-consumers perspective, and the processes used to access and transform the data. Three important findings for information systems professionals become available about solving *intrinsic*, *accessibility* and *contextual* data quality problems.

- Foster and Tuecke (2005) focused on previous technologies such as service oriented architecture, grid, on-demand, utility computing, software as service and others. All this to achieve a horizontal integration of the enterprise Information Technology.
- Youseff, et al., (2008) is an effort for a detailed ontology for the cloud computing environment, establishing the inter-relations between cloud components. Also presenting the ontology as a stack of layers, with its strengths and limitations, and its support on previous computing concepts.
- Korri, Taneli (2009) compares cloud computing with grid computing, and in the process describes the components of this new technology. He reviews the capacities of cloud computing, but also its weaknesses as the lack of standards and its impact with other cloud computing environments.
- Vaquero, et al., (2009) contributed with more than 20 definitions with the objective to get to a common ground, reviewing past work not only from a product view, but from the service view as well.
- Mikkilineni and Sarathy (2009) towards a cloud computing setting, they show a comparison of datacenters and the evolution of the Intelligent Network infrastructure in telecommunications. And they propose a next generation Virtualization Mediation Layer, to integrate server, network and storage virtualization, giving an internet platform to scale massively, while delivering reliable services.

Cloud computing has created many expectations as many discussions, since its quite new insertion in business (Rochwerger, B., et al., 2009). But prudence is required with new technology settings constructed from many previous technologies, since it usually provide a solutions to some previous problems, but it can also magnify others, or create new ones.
Researchers have developed studies on data quality and cloud computing separately, and on this study I propose the model of Input-Process-Output model, to validate the flow of the data sets through the emergent technology of cloud computing, where adaptations to processes are considered to take advantage of the new environment, so information quality attributes are improved, or reduce its negative effects. Then, an assessment method is utilized to demonstrate how data quality can be measured and evaluated in cloud computing.

Statement of the Problem.

Cloud computing certainly is a convergent evolution of a number of predecessor technologies, showing a solution for diverse challenges in the industry. For example, in 2006 two entrepreneurs launched a new dot.com called Animoto, having in April 2008 a peak on demand from 25,000 registered users to 250,000, and growing from a dozen of computers to nearly 5,000, in only three days (Smith, 2009). This exponential growth, would not have being possible, without a cloud computing solution.
The New York Times wanted to scan and place online images for a period of history covering 60 years; they were able to upload four terabytes in one day, with a cost of $25. This is possible with offerings like Amazon's Elastic Cloud Computing (EC2), which compare the monthly cost

of operating a basic server without any load, from $800 to $1,000, versus a rate from 10 cents to 15 cents an hour, for an equivalent computing power (Creeger, 2009).
But as stated by Geir Ramleth, CIO of Becthel and user of a cloud computing service “This is not a technology game, but a change-management game” (Creeger, 2009). And it is a place where a company such as Coghead could be started without buying any server, phones or software licenses, but relying completely on cloud services (Olsen 2006).
Important notes come from the data quality side, where errors in data can cost millions of dollars, but mainly it can discourage customers, and obstruct or even make impossible to implement new strategies (Redman, 1995). ”…50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous. The social and economic impact of poor-quality data costs billions of dollars” (Strong, et al., 1997).
In the same vein, “A leading computer industry information service firm indicated that it ‘expects most business process reengineering initiatives to fail through lack of attention to data quality’” (Wand and Wang, 1996).
As can be observed, the economic and business consequences imputable to a lack of information quality can be enormous and can also bring legal implications. This situation becomes worse with the addition of new technologies, which bring new options for business models. So, new tools aimed to measure are crucial, since all management information system need to be part of a management control system (Ackoff, R., 1967).

Statement of the Objective.

Various companies are joining the growing market of cloud computing, and each with a different number and type of services being used, and for which there are no standard definitions, but some terms used are: Software as a Service, Platform as a Service, Infrastructure as a Service, and others alike.
Although different studies have been made in the past about data quality issues, now, with the new paradigm of cloud computing, data quality problems need to be reviewed from this new perspective. Specifically, this study examined previous works around data quality dimension and attributes, considered the different points of view from academicians, who established the reasons and consequences of data quality problems. Then, I looked for the improvements and additional challenges, that a cloud computing environment brings to data quality problems. To achieve this, I used ten common problems associated to information quality attributes, identifying how the solutions previously found, are affected. Positive, negative or no change is expected from cloud computing onto the quality attributes, accordingly to the infrastructure utilized.
Since a new market is evolving, the functions and features worked in this study, will become the interest of many experts such as software developers, data architects, information systems resource managers, data managers, who will be interested not only on the evolution of cloud computing but its interaction with data quality.

Proposed Model and Taxonomy.

One of the fundamentals in Information Technology is the functional model of Input-Process-Output, also known as IPO+S model. Its simplicity helps us to trace the flow of data sets through processing environments, where an interaction could take place at the processing stage with other stored data (Redman, T., et al., 2000). At the end of the line, the output obtained is the information, which can be used as a new input, closing the feedback loop.

The study framework introduced in this research includes the data sets to be processed, the emergent technology of cloud computing formed among others, by virtualization, distributed computing, grid computing, web services and service oriented architecture (Zhang and Zhou, 2009). In the interaction with the processing of cloud computing, there are the common patches and solutions applied to data quality problems, represented here as the storage box. These patches and long-term solutions to usual data quality problems have been studied together with the dimensions of what define the quality of information (Strong, D.M., et al., 1997). Also on the output side, the information quality attributes are the objects to be verified, since their quality status will determine the overall information quality.

So the flow of the model goes from left to right, from the input of the data sets, to process of cloud computing. Here, the technology of cloud computing will be enhancing or worsening the patch and long-term solutions used to solve data quality problems; from here the importance of the interaction (Xu, M., et al. 2009). Finally, the outcome which will represent the information quality attributes, is the measure to adjust the process, using the feedback loop. This is considering the information system as an information manufacturing system, where the track and measure of data attributes can help to improve the quality of information (Ballou, D., et al., 1998).

Here we need to introduce the information quality attributes defined by researchers, and although it is not a standard definition, it has being used by many others in IT, and other disciplines like molecular biology (Naumann, F., et al., 1999). This study was based on the premise that in order to improve data quality, it is imperative to understand what the users of the data consider as quality. And the following table list the attributes grouped into four data quality categories (Wang and Strong, 1996):

Table 1. Information Quality Categories and Dimensions.

Information Quality	
IQ Category	**IQ Dimensions**
Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation
Accessibility IQ	Accessibility, Security
Contextual IQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of information
Representational IQ	Interpretability, Ease of Understanding, Concise representation, Consistent representation

Below, Figure 1 shows the IPO+S model, which is applied to a cloud computing environment in Figure 2. This comparison let us see not only the flow of the data, but the interactions of the information quality attributes, in a cloud computing setting, as an outcome of the process, and being at the same time the input of the next cycle, in the feedback loop.

Figure 1. Input-Process-Output classical model.

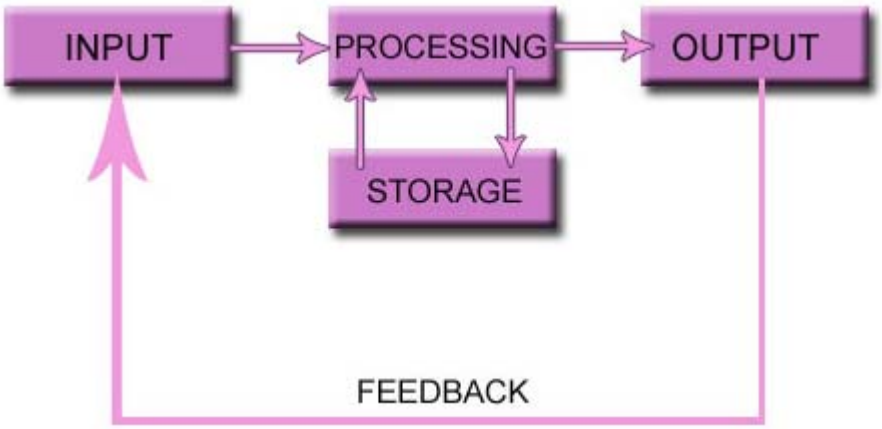
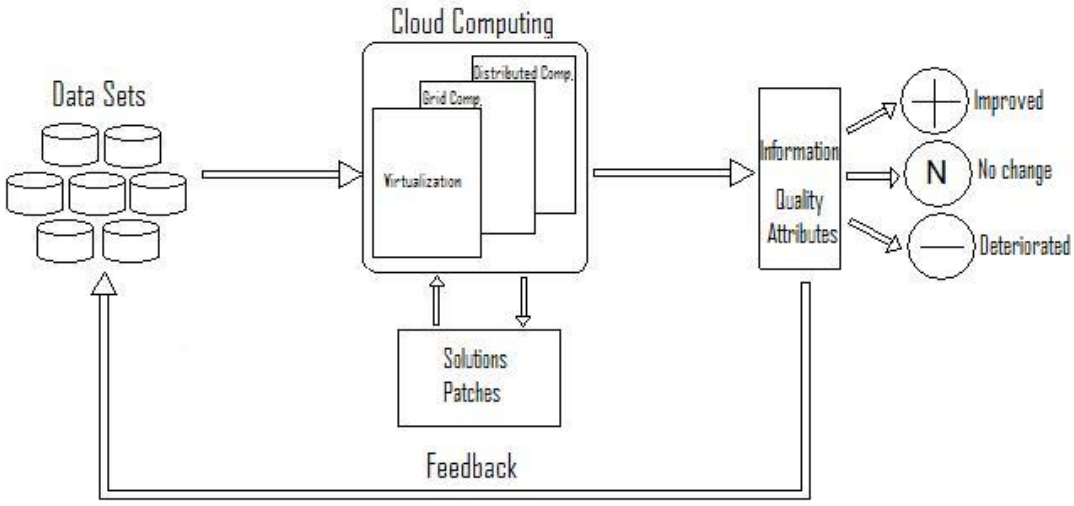


Figure 2. Information quality model within a cloud computing environment.



Considering this model we can establish if the new cloud computing paradigm improves or deteriorate the information quality attributes when the common problems arise in an information system. All these problems were studied by Strong, D., et al.(1997).

The first problem emerges when multiple sources of information, for example having multiple processes, generate different values for a piece of information, that should be the same.

Inconsistencies provoke that the believability, and by consequence, the value-added attributes are diminished. This is frequently find in companies, and with cloud computing it is expect to grow. Because of the ease to obtain a computing platform and the software as a service, each department, or division of the organization will be able to build their own local systems, without taking into account corporate standards. The long-term solution for this kind of problem is not directly affected by cloud computing, since it consists of a reexamination of information production process, in order to homogenize them.

Second, the information is produced using subjective judgments, having as a consequence different results to a common answer. Again, inconsistencies lead to the lost of information objectivity and believability. Cloud computing does not affect on the problem, nor to the solution to the problem, which is to work towards a continuous improvement, including additional training and better directions.

The third problem is when systemic errors are created when systems contain errors. These errors impact correctness, completeness and relevancy. The long-term solutions could include statistical process control being applied to information systems. Cloud computing does not provide any advantage or disadvantage on this point.

On the fourth issue, large volume of stored information makes it difficult to access in a reasonable time. The information quality dimensions touched here are the concise representation, timeliness, valued-added and accessibility. Cloud computing can contribute greatly, since it can provide not only computer resources as infrastructure required, or a specific platform for a certain system, but software as services, delivered through the internet. One temporary solution, or Patch, has been to extract subsets of information during the weekend in batch, so the information could be available for the following week. Now the extraction can be scheduled during the regular working hours, but utilizing the infrastructure from a cloud computing.

Next, we have as the fifth problem that distributed heterogeneous systems boost information inconsistency, not only by having diverse values, but different representations as well. The consistent representation, timeliness and value-added, are the dimensions impacted here. Cloud computing, since is based on the concept of distributed computing could encourage this issue, but the solution of developing a data warehouse populated in a centralized framework and solving inconsistencies, can be implemented on a bigger and best fit platform.

Continuing with the sixth problem, a nonnumeric data is hard to access due to the problem on how to index data types such as images, sound or video. Among the information quality dimensions affected are the concise representation, valued-added and accessibility. Cloud computing could provide the storage, but this is not the current problem, so there is no change here. The solutions are getting better as search and retrieval capabilities improve on the internet, and this can be applied to specific business needs.

As multimedia data growth and its storage is required, the seventh problem arises since there are issues with automated analysis for this kind of data. The information quality dimensions affected are consistent representation, relevance, and value-added. Cloud computing can again provide the infrastructure with bigger and faster platforms, but the solutions are coming with the progress of algorithm to analyze multimedia information like computed tomographies, podcast, etc.

Changes in consumer requirements always have been a concern for systems developers, which brings to us the eighth issue. The information quality impacted are relevance, value-added, and completeness. With the use of cloud computing an increase of users, and their demands for better and updated systems are expected to increase inevitably. And it does not contribute in positive or

negative way to the problem, nor to the solution which is change management, where the principal objective would be anticipated planning to changing consumer needs.
Arriving to the ninth problem, where the security is main topic, with privacy and confidentiality issues. This is confronted with the goal to facilitate the access of information. The information quality dimensions influenced are security, accessibility, and value-added. Seems as countermeasures one of the other, and cloud computing brings an scenario where this is increased, since it provides additional point of access, but the privacy and confidentiality are still concerns. Some solutions come from policies and procedures for the security, privacy and confidentiality of the data collected.
As the last problem considered, the tenth problem is the insufficiency of computational resources, which limits the access to systems. The information quality dimensions impacted are accessibility and value-added. The constant need to have updated equipment and bandwidth to speed up transactions, is easier and better achieved through cloud computing.
This model even though is simple, furnish us with the possibility to better trace the effects of the emerging technology of cloud computing towards the information quality attributes, and how it helps or not, to its solution. And at the same time, how it improves or deteriorates these attributes considering its offerings of infrastructure as a service, platform as a service, and software as a service.
Applications of this model will be useful as a base for information technology management, data management personnel, system development, project management, and data architecture.
The model is a diagrammatic representation of the different components of the data quality in a cloud computing setting, which corresponds to the components of an Input-Process/Storage-Output classical model. Here the parts interact and form the cycle of a data manufacturing processing (Ballou, D., et al.,1998), where data flow has an output, which in turn, will become the input in the following cycles, providing measures to improve the system. These kind of models have been studied previously (Ballou, D. and Pazer, H.L., 1985), but some limitation have evolve as technology moves and converge, as with cloud computing. This new environment brings new tools, but care is needed to obtain a benefit from its use.

Conclusion and implication for future research

A new environment, cloud computing, is gaining momentum, and with it, past problems need to be revisited, like information quality issues. We find that information quality attributes can be affected by new characteristics of cloud computing. The impact that cloud computing has over the information quality attributes, can be studied using an input-process-output diagram; where the flow of the data, and the interactions among the solutions for quality issues is shown. The new environment can effect positively, negatively, or have no impact at all, on the data quality attributes. Ten common problems related to information quality and their respective solutions, are verified to see how they are modified or not, being reinforced or thwarted.
New characteristics of cloud computing such as an enormous flexibility to provision the require infrastructure, platform for a specific applications, or software required, have huge implications for organizations economically and technically. This capacity to provision the required resources in a fraction of the time used just to do a basic installation of the equipment, and with a minimal

cost compared to an investment in equipment and services, offer new options for business models.

In spite of all the effort of previous researches to define and categorize the information quality attributes, and find long-term solutions to common quality issues of information systems, new technologies, and along with them new forms of interactions amongst people, generate new challenges.

Some of these new problems can be seen with exponential increase of semi-structured and unstructured data, with new datasets of multimedia. This creates new requirement for the storage of much bigger volume, but mainly a need to index, search and retrieve new kind of information. At the same time, cloud computing is emerging with new paradigms for infrastructure provisioning, platform and software settings.

But not all are promising news, since new technologies bring new opportunities and new problems. Cloud computing opens new environments, and with this openness possible problems with security, privacy and reliability arise. On the legal side, inquiries around control and ownership implications jump and shoot questions about the possibility of loosing the data in case of bankruptcy by the cloud computing provider, or legal obligations by the data holder to official requirements asking for access to it.

Finally, going back to the technical side, further research is required for applications requiring a really high reliability, with an availability of 99.9% or more. At the same vein, applications susceptible to latency need to verify ways to avoid it.

Cloud computing provides a flatter hierarchical environment, since it provides the abstraction for new infrastructure, platform or software needed. Along with this, new social interactions are developed, through social networking applications. Thus, a new “openness” data quality attribute needs to be considered, to measure how well the information meets the requirements of all these social-networking users. But additional freedom entails additional responsibilities, and here, security together with privacy and reliability are opposite to this openness.

References

Ackoff, R. (1967). Management Misinformation Systems. Management Science, Vol.14, No.4, December 1967.

Agmon, N. and Ahituv, N. (1987). Assessing Data Reliability in an Information System. Journal of Management Information Systems. Vol.4, No.2, Fall 1987.

Ballou, D., Wang, R., Pazer, H., Kumar.Tayi, G. (1998) Modeling Information Manufacturing Systems to Determine Information Product Quality. Management Science, Vol. 44, No. 4 (Apr., 1998), pp. 462-484

Ballou, D. and Pazer, H. (1985) Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. Management Science, Vol. 31, No. 2 (Feb., 1985).

Bezos, J. (2008) Jeff Bezos talks about Animoto at startup school 2008. YouTube. http://www.youtube.com/watch?v=uIc-VB-ke9o

Creeger, M. (2009) CTO Roundtable: Cloud Computing. Communications of the Association for Computing Machinery. August 2009, Vol.52, No.8.

Foster, I. and Tuecke, S. (2005). Describing the Elephant: The Different Faces of IT as Service. Association for Computing Machine. Queue *vol. 3, no. 6*

Foster, I., Zhao, Y., Raicu, I., Lu, S., (2008). Cloud Computing and Grid Computing 360-Degree Compared. Grid Computing Environments Workshop, 2008. GCE', 2008 - ieeexplore.ieee.org

Klems, M., Cohen, R., Kaplan, J., Gourlay, D., Gaw, P., Edwards, D., de Haaff, B., Sheynkman, K., Kepes, B., Sultan, O., Hartig, K., Pritzker, J., Doerksen, T., von Eicken, T., Wallis, P., Sheehan, M., Dodge, D., Ricadela, A., Martin, B., Berger, I.W., (2008). Twenty-one Experts Define Cloud Computing. Cloud Computing Journal. © 2008 SYS-CON Media Inc.

Korri, T. (2009) Cloud Computing: utility computing over the Internet. Helsinki University of Technology.

Levitin, A.V. and Redman, T.C. (1998). Data as a Resource: Properties, Implications, and Prescriptions. Sloan Management Review, Fall 1998.

Madnick, S. E., Lee, Y. W., Wang, R. Y., and Zhu H. (2009). Overview and framework for data and information quality research. ACM J. Data Inform. Quality 1, 1, Article 2 (June 2009) 22 pages. DOI = 10.1145/1515693.1516680. http://doi.acm.org/10.1145.1515693.1516680.

Madnick, S.E. and Wang, R.Y. (2009). Overview and Framework for Data and Information Quality Research. ACM Journal of Data and Information Quality, Vol. 1, No. 1, Article 2, Pub. date: June 2009.

Mikkilineni, R. and Sarathy V. (2009) Cloud Computing and the Lesson from the Past. 2009 18th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises.

Naumann, F., Leser, U., Freytag, J.C. (1999). Quality-driven Integration of Heterogeneous Information Systems. Humboldt-University of Berlin, Germany.

Olsen, G. (2006). Going Bedouin. Web Worker Daily. Sep 2006. http://webworkerdaily.com/2006/09/04/going-bedouin/

Pipino, L.L, Lee, Y.W., Wang, R.Y., (2002). Communications of the ACM, April 2002, Vol.45, No.4.

Redman, T.C. (1995). Improve Data Quality for Competitive Advantage. Sloan Management Review / Winter 1995.

Redman, T.C., Fox, C., Levitin, A., (2000). Data and Data Quality. Encyclopedia of Library and Information Science. Vol. 1.

Rochwerger, B., Breitgand, D., Levy, E., Galis, A., Nagin, K., Llorente, I., Montero, R., Wolsfthal, Y., Elmroth, E., Caceres, J., Ben-Yehuda, M., Emmerich, W., Galan, F. (2009). The RESERVOIR Model and Architecture for Open Federated Cloud Computing. IBM Journal of Research and Development, April 2009.

Smith, R. (2009). Computing in the Cloud. Industrial Research Institute, Inc. Sep-Oct.2009

Strong, D.M., Lee, Y.W., Wang, R.Y. (1997). Data Quality in Context. Communications of the Association for Computing Machinery. Vol.40, No.5, May 1997.

Strong, D.M., Lee, Y.W., Wang, R.Y. (1997). 10 Potholes in the Road to Information Quality. IEEE Computer Society, Vol.30, Issue 8, Aug, 1997.

Vaquero, L.M., Rodero-Medino, L., Caceres, J., Lindner, M. (2009) A Break in the Clouds: Towards a Cloud Definition. ACM SIGCOMM Computer Communication Review, volume 39, Number 1, January.

Wand, Y. and Wang, R.Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. Communications of the Association for Computing Machinery, Vol.39, No.11, November 1996.

Wei, J., Liu, L.C., Koong, K.S. (2006). An onion ring framework for developing and assessing mobile commerce security. International Journal of Mobile Communications, Vol.4, No.2, 2006.

Xu, K., Song, M., Zhang, X., Song, J., (2009) A Cloud Computing Platform Based on P2P. 2009 IEEE International Symposium on Information Technologies in Medicine and Education (ITME 2009), August 14-16, 2009, Ji'nan, P.R. China.

Xu, M., Cui, L., Wang, H., Bi, Y. (2009). A Multiple QoS Constrained Scheduling Strategy of Multiple Workflows for Cloud Computing. 2009 IEEE International Symposium on Parallel and Distributed Processing with Applications.

Youseff, L., Butrico, M., Da Silva, Dilma (2008). Toward a Unified Ontology of Cloud Computing. Grid Computing Environments Workshop, 2008. GCE'08. cs.ucsb.edu

Zhang, L. and Zhou, Q. (2009). CCOA: Cloud Computing Open Architecture. Web Services, 2009. ICWS 2009. IEEE International Conference on
6-10 July 2009 Page(s):607 – 616.