FROM QoD TO QoS
Data Quality Issues in Cloud Computing

Przemyslaw Pawluk, Marin Litoiu and Nick Cercone
York University, Toronto, ON, Canada

Keywords: Cloud computing, Quality of data (QoD), Quality of service (QoS), Quality of control (QoC).

Abstract: The concept of Quality of Data (QoD) has so far been neglected in the context of cloud computing. It was; however explored for the long time in the context of data exchange, data integration and information systems. Well established approaches like Total Data Quality Management, Data Warehouse Quality or Data Quality in Cooperative Information Systems have been proposed to calculate, store and maintain information about QoD. On the other hand concept of Quality of Service has been investigated in the context of Internet Systems, multimedia transmission and enterprise systems. It was also investigated in connection to cloud computing. The main goal of this work is to show direct connection between QoD and QoS. We show that assuring high QoD is necessary to achieve high QoS. We also identify major shortcomings of public cloud vendors in terms of provided configuration management data.

1 INTRODUCTION

Cloud computing refers to computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services. Cloud computing is a natural evolution of the widespread adoption of virtualization, service-oriented architecture, autonomic and utility computing (Vouk, 2008; Lim et al., 2009). Details are abstracted from end-users, who no longer have need for expertise in, or control over, the technology infrastructure "in the Cloud" that supports them. It does involve; however, certain level of control over virtual instances. This control requires high quality information about the system state. Virtual computing services becoming attractive for several reasons including adaptability, dynamic behavior and price. The Cloud computing leads to several research problems that have been of special interest. Here we will discuss provenance, which is analyzing history of data, trusted computing and automation of the Cloud control.

The Configuration Management Database (CMDB) provides a common trusted source for all IT data used by the business and promises to improve IT operational efficiency and increase alignment between the business and IT while reducing costs (EMA, 2008). The CMDB can be used also do support Cloud management.

Data quality problems occur along the entire data processing continuum. Data preparation is crucial and consists of several necessary operations such as cleaning data, normalizing, handling noisy, uncertain or untrustworthy information, handling missing values, transforming and coding data in such a way that it becomes suitable for the data mining process. Those methods are based on statistics and heuristics. The concept and importance of quality of data has been discussed many times in the literature (Ballou and Pazer, 1985; Batini and Scannapieco, 2006; Tupek, 2006; Wang and Strong, 1996) usually in context of the single data source. However, some research has been also done in the context of integrated data emphasizing the importance of data quality assurance in this context (Gertz and Schmitt, 1998; Naumann, 2002; Reddy and Wang, 1995). Data quality has been also considered in the context of data mining (Berti-Équille, 2007; Dasu and Johnson, 2003). As pointed by Beti-Équille (Berti-Equille and Moussouni, 2005) validity of results interpretation strongly relies on the data preparation process and on the quality of data set being analyzed. This is because methods such as data mining assume certain properties of data e.g. "nice" distribution.

Data quality problem has so far been neglected in the context of Cloud computing. Authors are not aware of any extensive work on this subject; however

Pawluk P., Litoiu M. and Cercone N..
FROM QoD TO QoS - Data Quality Issues in Cloud Computing.
DOI: 10.5220/0003558606970702
In *Proceedings of the 1st International Conference on Cloud Computing and Services Science* (IDQ-2011), pages 697-702
ISBN: 978-989-8425-52-2

some discussion have been found on Internet forums and blogs (Harzog, 2010; Vambenepe, 2010; Row, 2010). In those discussions some important aspects of data quality have been pointed.

In this work we discuss several data quality concerns and issues identified in the context of Cloud computing and CMDB. The remaining of this work has a following structure: Section 2 presents overview of research on Quality of Data, Section 3 describes the concept of Cloud Computing. In the Section 4 a link between QoD and QoS is presented. In the Section 5 current situation in public Clouds.

2 WHAT IS DATA QUALITY?

There can be found many different definitions of *Quality of Data* (data quality, QoD) in the literature. Researchers do not agree on one common definition of QoD and provide many essentially different definitions. This lack of common definition leads to defining QoD by providing dimensions – some better defined metrics that enables us to measure and compare some features of data sets. However this definition by definition failed since same dimension may be understand in a different way or same feature may be called differently by two researchers. This problem has been noticed by Wang and Strong (Wang and Strong, 1996). Commonly used definition (Tayi and Ballou, 1998; Wang and Strong, 1996; Orr, 1998) defines quality as "fitness for use". It implies the relative nature of the quality concept. As stated in (Orr, 1998) understanding of quality depends strongly on how users actually use the data in the system, since they are ultimate judges of the quality. There is no common standard of QoD, however we can find an ISO document *ISO8402:1995 Quality Management and Quality Assurance Vocabulary* (ISO, 1994) or its newer version *ISO9000:2005 Quality management systems – Fundamentals and vocabulary* (ISO, 2005). It provides a formal definition of quality as: "The totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs" (ISO, 1994). It is clear that the main authority in terms of QoD is the user and his requirements are main guidelines to define and measure QoD.

2.1 Data Quality Dimensions

Data quality is defined often through quality dimensions (called sometimes quality factors). We will use words dimension and factor interchangeably in the remainder of this work. There is over hundred different dimensions identified in some publications (Wang and Strong, 1996). We do not discuss all QoD factors in this work. We rather concentrate on those factors that can be applied into the context of Quality of Service. A comprehensive discussion of different quality dimensions can be found in (Wang and Strong, 1996) and (Batini and Scannapieco, 2006).

2.1.1 Data Decay – Time Related Factors

There is a subset of quality factors directly correlated with time. This characteristic is intuitive and we can easily point such factors. The only problem is that there are different meanings of time-related terms proposed in the literature. For example Naumann (Naumann, 2002) defines timeliness as the average age of data in source. Timeliness in other sources "refers to the length of time between the reference period of the information and when we deliver the data product to our customers" (Tupek, 2006).

Segev (Segev and Fang, 1990) defines currency as the time interval between extraction and delivery. The currency in this form has been named timeliness by Wang (Wang and Strong, 1996). This definition in our opinion is best fitted for the Cloud systems. It may seen as a delay between consecutive readings of a Cloud state.

2.1.2 Accuracy

Accuracy is included by most data quality studies as a key factor (Parssian et al., 2002; Batini and Scannapieco, 2006; Wang et al., 2005; Ballou and Pazer, 1985; Gertz and Schmitt, 1998). Although the term has an intuitive appeal, there is no commonly accepted definition of what it means exactly (Wand and Wang, 1996). Ballou and Pazer (Ballou and Pazer, 1985) describe accuracy as "the recorded value is in conformity with the actual value." Kriebel (Kriebel and Moore, 1982) characterizes accuracy as "the correctness of the output information." Thus, accuracy in this case appears as the term viewed as equivalent to correctness.

In (Batini and Scannapieco, 2006) accuracy is defined as "the closeness between a value v and a value v', considered as the correct representation of the real-life phenomenon that v aims to represent." The simple example can be the name of the city $'Toronto'$, the value $v = 'Tronto'$ is incorrect (inaccurate) and $v' = 'Toronto'$ is correct (accurate).

Accuracy can be also seen as an "error bar". In other words an error of the measurement. In case of Cloud computing, managing is based on aggregated values of readings. Accuracy can be expressed in such case as a standard deviation in the sample. It is only a suggestion and this problem requires further investi-

gation. We do not discuss this issue in this paper.

2.1.3 Completeness

Term defined by Naumann (Naumann, 2002) is coinciding with nullability and is "the quotient of the number of non-null values in a source and the size of the universal relation." It means that the less null values in the relation, the higher value of completeness is (more complete is the relation).

In (Wang and Strong, 1996) the completeness is defined as "the extent to which data are of sufficient breadth, depth, and scope for the task at hand." Bobrowski (Bobrowski et al., 1998) claims that it expresses that every fact of the real world is represented in the information system.

The problem of completeness definition in the context of Cloud system requires further investigation. We have to define, what does it mean that our system representation (variable set) is complete.

3 CLOUD COMPUTING

Cloud computing refers to computation, software, data access, and storage services that do not require end-user knowledge of the physical location and configuration of the system that delivers the services. Cloud computing is a next step in the evolution of the widespread adoption of virtualization, service-oriented architecture, autonomic and utility computing (Vouk, 2008; Lim et al., 2009). Details are abstracted from end-users, who no longer have need for expertise in, or control over, the technology infrastructure "in the Cloud" that supports them. Virtual computing services becoming attractive for several reasons including adaptability, dynamic behavior and price. The Cloud computing leads to several research problems that have been of special interest. Here we will discuss different dimensions of QoD.

Cloud computing uses remote virtual servers for storage and all processing of data. Data quality, therefore, becomes one of primary requirements and administrators should address this aspect before deciding on a Cloud computing vendor. Lately, only *Quality of Service* (QoS) was considered for Clouds. QoS refers to a broad collection of networking technologies and techniques. The goal of QoS is to provide guarantees on the ability of a network to deliver predictable results. Elements of network performance within the scope of QoS often include availability (uptime), bandwidth (throughput), latency (delay), and error rate. Clouds fall into one of following three types of system(Vaquero et al., 2008):

- *Software as a Service* that is defined as a provider supplying remotely run software packages on a utility based pricing model. (e.g. online text editors or spread sheets)
- *Platform as a Service* that is defined as a provider offering additional layer of abstraction above the virtual infrastructure. PaaS offers built-in scalability traded off by some restrictions of software that can be deployed.
- *Infrastructure as a Service* that is defined as a provider provisioning compute and storage resource capacity through virtualization. IaaS allows physical resources to be assigned and split in dynamic manner.

Three types of Cloud systems form layers in a sense that higher layer can deploy and utilize the lower level features (Armstrong and Djemame, 2009).

4 QoD AND QoS

Quality of data can be considered on two distinct levels in Cloud computing environment. The first level, probably most obvious, is quality of data deployed into the Cloud (customers data). It has to be processed in consistent way. Data quality indicates the degree of excellence within the data, its state of completeness, validity, and accuracy that enables it to perform further functions. This in turn, enables the user to obtain the necessary information required for operational reasons or to assist in decision making and planning. Data of high quality produces results that need to be reliable and correct. In essence, if you choose Cloud computing, data quality needs to be accurate and in reliable formats. Ideally the Cloud infrastructure should not interfere with data on this level and, if it is not explicitly required, leave quality assurance to customers.

The other level of quality in Cloud is quality of internal data such as configuration management records (CMR) or simply measurements of resources usage. It is perceived as meta-data describing the Cloud system. Working on that level, the goal is to assure high QoS. We will see that it can be done through high QoD. Let us consider now how QoD and QoS interfere in the context of Cloud computing. To do that we have to analyze important aspects of QoS, techniques allowing to achieve them and identify how QoD impacts those techniques.

4.1 Dimensions of QoS

In characterizing the QoS of activities, it is necessary to identify dimensions along which QoS can be measured and quantified. In this work we consider QoS from the perspective of a service provider. The meaning of enlisted dimensions may change when consider from other actors (e.g. end-user) point of view. It is useful to group sets of QoS dimensions into QoS categories where each category contains dimensions pertaining to some logically identifiable aspect of QoS. Campbell (Campbell, 1996) distinguished following categories:

- *system reliability* – contains system-related reliability dimensions (e.g. MTBF, MTTR)
- *timeliness* – contains dimensions relating to the end-to-end delay of data flow
- *volume* – contains dimensions that refer to the throughput of data in a flow
- *criticality* – relates to the assignment of relative priority levels between activities
- *quality of perception* – is concerned with dimensions such as screen resolution or sound quality
- *cost* – understand as a fee paid by a service provider to the Cloud vendor

Let's now take a closer look at those categories and see if they depend upon data quality. In some cases such dependency seems to be obvious in other cases it requires deeper investigation and is not visible at first glance.

4.1.1 Timeliness

Timeliness category contains dimensions relating to the end-to-end delay of data flow. Such delay depends upon several aspects of the system deployed in the Cloud. One of those factors is number of running instances. This number can be changed based on information about system load and resources usage provided by Cloud vendor. If such information is not fresh (up-to-date) it is impossible to take just-in-time decisions. Delayed decisions can easily lead to lower QoS. In particular the number of active instances may not be sufficient to meet assumed response time and to satisfy end-users' expectations.

4.1.2 Cost

Cost category refers to the cost of processing and usage. Cost in public Cloud depends, among many other factors, directly on the number of active instances. This number can be decreased as load is lower. Clearly, freshness (currency, timeliness) of information about load provided by Cloud vendor strongly impacts the ability to undertake valid decisions.

Checking Amazon EC2 pricing we can see there that a price for default Windows instance is \$0.12 per hour. Default policy says also that "pricing is per instance-hour consumed for each instance, from the time an instance is launched until it is terminated. Each partial instance-hour consumed will be billed as a full hour" (Amazon, 2011a). Amazon EC2 offers monitoring with a time window of 30 minutes. How does it influence cost of computing? Let's analyze following case. An application is deployed on Amazon EC2 public Cloud and uses up to three instances. Additional instances are launched when request rate reaches following thresholds: 4,000 for the second instance and 7,000 for the second instance. Instances are terminated when request rate is lower than 6,000 and 3,000 respectively for third and second instance. If such situation repeats every day it means we loose $30 * \$0.12 = \$3.6$ every month, or about \$44 each year. This calculation is done for the smallest instance offered by Amazon EC2 and cost can be higher depending on the configuration (for high-CPU on-demand instances this cost can be as high as \$434 per year in the same scenario).

On the other hand, there is a certain cost of each message (measurement) sent. In case of Amazon EC2 it is \$0.008 per message. It does not seem to be much; however measuring only one variable hourly gives us about \$70 per year. Increasing the frequency and measuring every thirty minutes gives \$140 yearly. In such case there is a trade-off that needs to be made based on dynamism of the system.

There are of course certain methods to act proactively in such situation. One can analyze trend to predict the time when certain instance should be terminated, however even for those methods up-to-date information is necessary for accurate predictions. The accuracy impacts then cost and/or performance of the system.

4.1.3 Other Dimension Category

System Reliability category contains system-related reliability dimensions such that Min Time Between Failures (MTBF) or Min Time To Recovery (MTTR). *Volume* category contains dimensions that refer to the throughput of data in a flow. *Criticality* category relates to the assignment of relative priority levels between activities. *Quality of Perception* is a category that is concerned with dimensions such as screen resolution or sound quality and refers to user perception. More detailed discussion of QoS dimension can be found in ISO/EIC "Information technology – Quality

of service: Framework" standard (ISO/IEC, 1998).

4.2 Why it is Important?

We have shown that certain dimensions of quality of data has significant impact over some dimensions of quality of service. At this point we would like to summarize our point of view.

We have shown examples of influence of certain dimensions of QoD such as freshness and accuracy on some dimensions of QoS. We claim that the interconnection between QoD and QoS can be utilized to improve *Quality of Control* (QoC) (Marti et al., 2002). This metric allows us to measure how good and how fast the system can react on certain events. For example, how fast new instance can be started to handle excessive number of request. Because of lack of space, we do not discuss this concept in detail.

5 HOW DOES IT WORK IN PUBLIC CLOUDS?

In this section we present current landscape of the public Cloud market. Our goal is to show what is currently provided by Cloud vendors. There are four main commercial providers of Cloud services on the market.

Amazon was the first company supplying Cloud infrastructure early in 2006. *Amazon Web Service* (Amazon, 2011b) provides PaaS on pay per use basis. They provide two products the *Amazon Elastic Compute Cloud* (EC2) and the *Amazon Simple Storage Service* (Amazon S3). Amazon provides also set of API's. In pay-per-use model Amazon is charging per the time the instance is active. Additional cost applies for messages (state of the system), storage etc.

Another provider of Cloud services is Google. Google provides SaaS through *Google Apps* (Google, 2011b) software and an PaaS via *Google App Engine*(Google, 2011a). The Google App Engine provides the architecture that Google Apps runs on. They also use pay per use economical model charging service provider per application and per user.

IBM provides PaaS based on API's created by Amazon. It is known as *IBM's Research Compute Cloud*(IBM, 2011a). IBM provides also *IBM Computing on Demand* (IBM, 2011b) that are addressed to supply enterprise Cloud Computing. IBM uses economical model similar to the model used by Amazon and charges per hour of usage. Prices varies depending on operating system and virtual server configuration. Microsoft is not providing Cloud services. The company is, however, developing the *Azure Service Platform*(Azure, 2011). Azure is PaaS operating system that incorporates many Microsoft's packages. It can be utilized by licensed Cloud vendors as all-in-one Cloud software solution. In this case charges are also calculate per hour using pay-per-use model.

6 CONCLUSIONS AND FUTURE WORK

In this work we pointed important data quality issues arising in the area of Cloud computing and their effects for certain dimensions of QoS. This correlation of QoS and QoD requires deeper investigation. It is clear; however that data quality assurance is necessary to achieve high quality of service.

We have shown here the intuitive examples of correlation between different dimensions of QoS and QoD such as cost (QoS) and freshness (QoD), timeliness (QoS) and freshness (QoD). We are going to investigate this issue in depth.

This work shows the interconnection between QoS and QoD dimensions in informal way and does not provide quantitative methods of assessment. Our future work will be concentrated on formalizing this connection and providing quantitative methods of its assessment. We want to map data quality dimensions into quality of service dimensions and design functions modeling those mappings in mathematical way.

Our long term goal is to provide a model combining quality of data and quality of service, and improving at the same time quality of control by enabling just-in-time decisions and reducing settling time. To achieve this goal, we are going to develop a new, quality-aware type of autonomic manager. This can be achieved by development of quality-aware sensors and effectors.

Our first step will be experimental evaluation of sensors adjusting the measurement interval dynamically depending on the change rate of the measured value. Intuitively, rapidly changing values should requires more frequent measurement. Dynamic adjustment of the interval is expected to optimize certain dimensions of QoS and the cost (or overhead generated by frequent measurement) at the same time.

REFERENCES

Amazon (2011a). Amazon ec2. http://aws.amazon.com/ec2/.

Amazon (2011b). Amazon web service. http://aws.amazon.com.

Armstrong, D. and Djemame, K. (2009). Towards quality of service in the cloud. In *Proceedings of 25th UK Performance Engineering Workshop*.

Azure (2011). Azure. http://www.microsoft.com/azure.

Ballou, D. and Pazer, H. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2):150–162.

Batini, C. and Scannapieco, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Berti-Équille, L. (2007). Data quality awareness: a case study for cost optimal association rule mining. *Knowl. Inf. Syst.*, 11:191–215.

Berti-Equille, L. and Moussouni, F. (2005). Quality-Aware Integration and Warehousing of Genomic Data. In *Proceedings of the 2005 International Conference on Information Quality*.

Bobrowski, M., Marr, M., and Yankelevich, D. (1998). A software engineering view of data quality. In *European Quality Week Conference*.

Campbell, A. T. (1996). *A Quality of Service Architecture*. PhD thesis, Lancaster University.

Dasu, T. and Johnson, T. (2003). *Exploratory Data Mining and Data Cleaning*. Wiley-Interscience.

EMA (2008). How to define detailed requirements for your enterprise cmdb project: A hands-on workbook.

Gertz, M. and Schmitt, I. (1998). Data Integration Techniques based on Data Quality Aspects. In Schmitt, I., Türker, C., Hildebrandt, E., and Höding, M., editors, *Proceedings 3. Workshop "Föderierte Datenbanken", Magdeburg, 10./11. Dezember 1998*, pages 1–19. Shaker Verlag, Aachen.

Google (2011a). Google app engine. http://code.google.com/appengine.

Google (2011b). Google apps. http://www.google.com/apps/business.

Harzog, B. (2010). Is the cmdb irrelevant in a virtual and cloud based world? Blog entry: http://www.virtualizationpractice.com/blog/?p=5726.

IBM (2011a). Ibm cloud computing. http://www-935.ibm.com/services/us/cloud/index.html.

IBM (2011b). Ibm computing on demand. http://www-03.ibm.com/systems/deepcomputing/cod/.

ISO (1994). *ISO 8402 Quality Management and Quality Assurance: Vocabulary.* ISO. Withdrawn standard.

ISO (2005). *ISO 9000:2005 Quality management systems – Fundamentals and vocabulary.* ISO. Published standard.

ISO/IEC (1998). *ISO/IEC 13236:1998. Information technology – Quality of service: Framework*. ISO/IEC.

Kriebel, C. H. and Moore, J. H. (1982). Economics and management information systems. *SIGMIS Database*, 14(1):30–40.

Lim, H. C., Babu, S., Chase, J. S., and Parekh, S. S. (2009). Automated control in cloud computing: challenges and opportunities. In *Proceedings of the 1st workshop on Automated control for datacenters and clouds*, ACDC '09, pages 13–18, New York, NY, USA. ACM.

Marti, P., Fuertes, J. M., and Fohler, G. (2002). Improving quality-of-control using flexible timing constraints: Metric and scheduling issues. In *In IEEE RTSS*.

Naumann, F. (2002). *Quality-driven query answering for integrated information systems*. Springer-Verlag New York, Inc., New York, NY, USA.

Orr, K. (1998). Data quality and system theory. *Commun. ACM*, 41(2):66–71.

Parssian, A., Sarkar, S., and Jacob, V. S. (2002). Assessing information quality for the composite relational operation join. In *IQ*, pages 225–237.

Reddy, M. P. and Wang, R. Y. (1995). Estimating data accuracy in a federated database environment. In *CISMOD*, pages 115–134.

Row, J. R. (2010). All about cloud computing and data quality. *http://www.brighthub.com*.

Segev, A. and Fang, W. (1990). Currency-based updates to distributed materialized views. In *Proceedings of the Sixth International Conference on Data Engineering*, pages 512–520, Washington, DC, USA. IEEE Computer Society.

Tayi, G. K. and Ballou, D. P. (1998). Examining data quality. *Commun. ACM*, 41(2):54–57.

Tupek, A. R. (2006). Definition of data quality.

Vambenepe, W. (2010). Cmdb in the cloud: not your fathers cmdb. Blog entry: http://stage.vambenepe.com/archives/1527.

Vaquero, L. M., Rodero-Merino, L., Caceres, J., and Lindner, M. (2008). A break in the clouds: towards a cloud definition. *SIGCOMM Comput. Commun. Rev.*, 39:50–55.

Vouk, M. A. (2008). Cloud computing issues, research and implementations. *ITI 2008 30th International Conference on Information Technology Interfaces*, 16(4):31–40.

Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Commun. ACM*, 39(11):86–95.

Wang, R. Y., Pierce, E. M., and Madnick, S. E. (2005). *Information quality*, volume 1 of *Advances in management information systems: Information Quality*. M.E. Sharpe.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33.