DATA INTEGRATION THROUGH THE CLOUD
How to Combine Internal and External Data Sources – A Design Study

Patrik Hitzelberger, Paulo da Silva Carvalho and Fernand Feltz
Centre de Recherche Public Gabriel Lippmann, 41 rue du Brill, L-4422 Belvaux, Luxembourg

Keywords: Data Integration, Open Data, Cloud Computing.

Abstract: This short paper focuses on the application of cloud computing principles and solutions to the domain of data integration. After an introduction to the topic, data integration is shortly discussed, and some quality criteria for data integration solutions, including infrastructure and the organizational context, are presented. Afterwards, cloud computing and possible cloud-based data integration scenarios are discussed. The before-mentioned quality criteria are revisited especially relative to public cloud deployment scenarios. Finally, a design study for the examination of cloud-based data integration that focuses on open data integration for an environmental data management application is proposed.

1 INTRODUCTION

The integration of information systems is a fundamental task for software architects and developers. Academia and industry have been working in this field for at least two decades now (cf. e.g. Hasselbring, 2000). This has resulted in a plethora of bespoken software architecture patterns (Berbner et al., 2005) and products for information systems integration (Bernstein and Haas, 2008) that cover certain aspects of integration.

This short paper focuses on data integration as one of these aspects, and examines opportunities and challenges of the cloud paradigm applied to this field. We will argue that cloud-based data integration is probably not a panacea for all integration needs, and to be scrutinized is which technical architecture fits best for an organization and its requirements.

The paper consists of two main parts. We will first analyse briefly the state-of-the-art of data integration and data integration in the cloud, and examine some ongoing work. In the second part, we present research questions and our recently started research project in the domain of cloud-supported data integration with a focus on open data (Miller et al. 2008) integration.

2 DATA INTEGRATION

Lenzerini (2002) defines data integration as "the problem of combining data residing at different sources, and providing the user with a unified view of these data".

Hasselbring (2000) has identified three dimensions for describing this problem of the integration: distribution, autonomy and heterogeneity of the underlying technical systems, organisations and data. Data to be integrated can exist on geographically distributed systems run by different organisations, in different formats.

Advanced data integration technologies and tools that tackle the inherent integration problems are a prerequisite for application domains like BI, CRM and Master Data Management. The focus of such tools used to be on read-only applications, meaning that the integrated data view does not change the data. Read-and-write scenarios, where integrated data is also modified, seem to become more important however (Yahanna and Gilpin, 2012).

Data integration concerns more and more electronic information that is available and shared on the Internet, coming from different origins and sources: private companies (e.g. annual reports), public entities and governments. The concerned data is either an implicit result of the growth of the Internet, or explicitly fostered by open data and similar initiatives that aim at using the Internet for public data access (Dekkers et al., 2006).

There are different scenarios for the combination of organisation internal data with external, public or non-public sources. It can e.g. foster or permit the cooperation between enterprises, government

Hitzelberger P., da Silva Carvalho P. and Feltz F..
DATA INTEGRATION THROUGH THE CLOUD - How to Combine Internal and External Data Sources – A Design Study.
DOI: 10.5220/0003961001770182
In *Proceedings of the 2nd International Conference on Cloud Computing and Services Science* (CLOSER-2012), pages 177-182
ISBN: 978-989-8565-05-1

agencies and other private or public entities (Halevy et al., 2006). In unilateral scenarios, organisations can integrate external data in order to improve internal analytical or reporting processes. Other projects focus on the external integration of open data for public re-use, and tackle the paramount challenge of integrating (semi-) structured, but semantically different data (Böhm et al., 2010). We focus exemplarily on the application of data integration in environmental research, where the integration of internal and external data, becomes more and more important.

The original objective of data integration is to generate data sets with new, added value, by homogenising and cleaning the original data sources. (Halevy et al., 2006). Cloud computing extends this idea by virtualizing the necessary IT infrastructure. The general advantages of such a virtualization are known (Al-Zoube, 2009). We focus in the next paragraphs on some of the goals and quality criteria for data-centric integration solutions. These criteria are based on the assumption that data is being regarded as an asset by organisations that have to report a "single version of the truth" (Khatri and Brown, 2010). Industry markets their own data integration solutions also known as *Information as a Service* and *Data Virtualization*, with numerous applications in all sectors. In this paper, we stick to the notion of data integration, because all terms converge on the integration of heterogeneous data sources.

Theoretically, the data from these sources can be copied, migrated or accessed on-line. If external data is integrated, it might be inevitable to persist and copy this data, because it is either required to snapshot the actual state of data that changes over time, or because the (long-term) online availability of external sources cannot be guaranteed. The more external sources are integrated, the more this issue becomes relevant. Given that, there are obviously integration scenarios where scalable data persistence mechanisms are indispensible for many or all data sources, and online-access and integration at runtime is not sufficient. Cloud computing seems to offer this.

In the next paragraphs, we will present some important criteria for evaluating data integration solutions. The list is not exhaustive, but tries to present some key issues that we will examine in our further work.

2.1 Data Quality

There is rich literature about information and data quality models and criteria. Naumann and Rolker (2000) have already underpinned that "Information quality (IQ) is one of the most important aspects of information integration on the Internet" (which is only partly our objective, but applies to any integration scenario). They identified 22 information quality criteria, classified into subject (e.g. *relevancy*, *comprehensibility*), object (e.g. *completeness*, *price*) and process criteria (e.g. *accuracy*, *availability*). Khatri and Brown (2010) ,who are less exhaustive and more focused on organisational data governance enumerates *accuracy*, *timeliness*, *completeness*, and *credibility*.

It is obvious that there is an impact of data source quality, heterogeneity and the number of sources to integrate on the resulting integration solutions. Furthermore, the criteria might be different for the different sources.

2.2 Technical Infrastructure

Data can only be accessed on sound information infrastructures. As for data quality, there are different metrics and requirements for assessing IT infrastructures, like

- *reliability*, with criteria like fail-safety and redundancy, backup and long-term archiving features and so on
- *performance* and *scalability* of hardware, software and network infrastructure elements

Data integration solutions require specific IT infrastructures. They can e.g. produce enormous data volumes to manage, resulting in extended requirements for storage, backup and network capacities.

A detailed discussion of all related IT infrastructure standards like e.g. ISO 27001, ITIL and the related literature is outside the scope of this short paper.

2.3 Process and Compliance

The quality of IT solutions is a result of technical, but also procedural and organisational measures. Quality criteria can only be defined relative to domain and application specific requirements.

Regarding the specifically data integration related requirements, compliance and legal frameworks can define, which data has to be stored and managed in what way, e.g. with respect to data retention times, data life-cycles, data location, and data protection provisions to respect. Reporting obligations become more and more important in

sectors like finance and health (Anderson et al., 2011).

The necessary measures, processes and organisational decisions in order to reach and maintain compliance to the sector specific frameworks can be denoted as *Data Governance* (Khatri and Brown, 2010). Replacing crucial parts of existing data integration solutions has a chief impact on it.

2.4 Costs

Naumann and Rolker (2000) define *price* as a data quality criterion (see above). Any information system technology assessment must take into account and compare investment and operational costs of the solutions. Given that costs reduction is one of the central arguments in the marketing for cloud offers, research must focus on costs and the related payment models and their potential impact on data integration. The objective is to evaluate cloud based data integrations in comparison to conventional solutions.

2.5 Other Criteria

Quality and assessment criteria for data integration solutions can be detailed further, as mentioned above. Process-related criteria like speed and ease of development, maintenance costs etc. (D'Agostino et al., 2010) should be taken into account. If the focus of the solution is external collaboration and/or data integration, the weighting of the criteria changes considerably (Doelitzscher et al., 2011).

In the next chapter we discuss some of the presented criteria with respect to cloud data integration solutions.

3 CLOUD DATA INTEGRATION

3.1 Cloud Technologies

There are many definitions of cloud computing and cloud technologies. We follow Baars and Kemper (2010) by abstracting from the details and looking at cloud computing as “a distributed, net-based architecture where resources can be dynamically rearranged”. It seems commonly agreed that

- the technical access to this architecture is service-oriented,
- infrastructure, platform and software (IaaS, PaaS and SaaS) are the main layers for accessing it,
- public and private clouds are the primary deployment models. Hybrid and community forms are possible (Mell and Grance, 2011),
- payment and scalability is demand-oriented.

Technology-wise, cloud solutions offer potential for

- a complete virtualization of data integration by migrating all relevant data sources and infrastructure to the cloud
- mixed scenarios where only parts and/or copies of the internal data to integrate is put into the cloud, and an on-premises infrastructure is kept,
- on-premises integration where cloud technologies are used in a private cloud. In such deployment models, data does not leave the organisation.

Some available SaaS applications address typical data-integration driven domains, like BI or CRM, offered either in public or private clouds. They offer supporting tools and standard connectors for migration of and/or the interfacing to internal applications, but are restricted to specific application domains and organisational data. There are also more versatile tools that are marketed as universal data integration platforms, coming from well-established business players in this domain.

It seems that independently from buy-or-build decisions when investing in the cloud, the crucial data-integration related issue is related to the chosen deployment model. If public or hybrid cloud models are part of the solution, some or all data has to leave the concerned organisation(s), either as copy or after migration. Initially, this *disintegrates* existing IT processes and infrastructures. In-house private cloud solutions seem to be less disintegrating, and are considered by some authors as possible transitory solutions for public cloud based solutions later on (Géczy et al., 2012).

Both scenarios require the development of service-oriented accesses to the integrated data. This results in a “Data-as-a-service” view on data assets.

3.2 Data Integration by and in the Cloud

In the following, we discuss briefly the general criteria and characteristics of data integration that have been introduced in chapter 2. We argue that it is vital to understand the potential and risks using cloud computing for data integration. Generally, literature confirms that it is at least questionable if every organization should move all or parts of their

data assets into the cloud. Kim (2009) asks e.g.:

- Which information must be moved?
- Which information cannot be moved?
- What is the availability of that kind of system?
- Is the system performance affected?
- What is the level of security of that kind of system? Is it trustworthy?

The next subparagraphs sketch some of these issues.

3.2.1 Data Quality

A cloud computing approach does not modify the requirements for data quality. There are proven and mature tools that permit data cleansing for duplicate detection, data fusion etc. (Halevy et al., 2006). As soon as external data sources are integrated, data quality and lineage become more difficult to attain.

Data access tools and layers have to be redesigned and adapted to new languages and service-oriented methods, and that data-models might change.

3.2.2 Technical Infrastructure

One of the central arguments of cloud providers is the fact that clouds offer the dematerialization of the management of data assets by replacing complex and expensive internal infrastructures by on-demand cloud solutions. This is the case for migration solutions into public clouds.

Especially for solutions with high volumes of data to integrate, organizations must be aware of the fact that the current public cloud solutions depend on the public internet as transport layer. Current internal SAN solutions for storage reach 1-10GB/s guaranteed network performance, which is by orders of magnitude faster than the vast majority of internet connections available, especially for small and medium enterprises and organisations.

Also, it seems to be difficult, based on the available public cloud providers SLA models, to define a clear level of reliability for end-to-end processes when cloud infrastructure elements are part of the solutions (Géczy et al., 2012). Recently, data loss and outages of large public cloud providers have been reported (Blodget, 2011).

3.2.3 Process and Compliance

The externalization of data to public cloud requires to adapt (or re-integrate) elementary parts of an existing IT infrastructure and the related organizational and process frameworks. Accountability and compliance to legal and sector-specific frameworks might have implicit or explicit provisions that hamper or circumvent cloud projects.

This is mainly due to the fact that public cloud solutions represent two fundamental modifications of existing on-premises solutions (Adkinson-Orellana et al., 2011):

- Data is moved or copied outside the organisation
- Data is managed or stored by a third party

In contrast to conventional outsourcing solutions where dedicated resources are run and managed by third parties, the business model of cloud providers is by definition focused on their internal economy of scales; meaning that data location and sharing behind the service-oriented data access level is

Table 1: Comparison of public cloud and on-premises data integration characteristics.

	Public Clouds	On-premises (incl. private clouds)
Data quality	Same restrictions and technologies apply. Possible impact of new data access methods	Proven technologies and tools available. Inhouse development, COTS solutions.
Technical Infrastr.	Desintegration or migration of own infrastructure – public network and vendor dependency. Potential bandwith issue.	Internal infrastructures with definable reliability and characteristics.
Data managemt.	External vendor must be integrated in compliance and legal context, trust model changes	Internally managed procedures and accountability. Data stays in organisation boundaries.
Costs	Potential cost-savings through on-demand cost-models. Economy of scales at provider side. (Armbrust et al., 2009). Integration costs.	Traditional IT cost models (own, leased, outsourced soft- and hardware costs). Maintenance and development costs.
Online collaborat. and sharing	Collaboration and sharing within the external cloud over well-defined services	Bespoken bilateral information exchange with infrastructure on both sides
Speed and ease of deployment	Standardized products, quick on-demand setup of test and staging servers, etc. (D'Agostino et al., 2010)	Depends on buy-or-build decisions and customization effort.

hidden from their customers. Physical data locations, the number of copies, backup strategies and so on are normally not part of SLAs for clouds. In terms of legal requirements, this can lead to conflicts of law, because “data may move from one jurisdiction into another in milliseconds” (Spies, 2011). In some countries it is explicitly forbidden to export certain kinds of data.

Table 1 summarizes the discussed issues with data integration and compares them to on-premises strategies, including private clouds.

4 DESIGN STUDY: CLOUD INTEGRATING OPEN DATA

Based on the discussed literature and practical experience, we argue that the migration of data integration applications into (in particular: public) clouds disintegrates existing information system architectures. The redesign of a new system cannot be without costs, and might even be impossible in a given setting. The risks and costs must be balanced against the potential advantages of a comprehensive data virtualization as detailed above.

In order to judge and evaluate cloud data integration solutions and offers, we will conduct a design study (Hevner et al., 2004) that will try to find answers to the following principal challenges:

- How to design and build a generalized cloud data virtualization application that can integrate internal organization data and external open data?
- What are the necessary technical and organizational prerequisites?
- How to re-integrate and adapt existing data integration architectures?
- Can private cloud integration solutions serve as a transitional solution for public cloud data integration solutions later on?

Based on the methodological approach and these questions, the main design artefact will be a software prototype with the following preliminary global specification:

- migration of an existing database application for soil data management into a private cloud solution based on open source software
- Integration of available open environmental data into this cloud
- Adaption and/or redevelopment of the existing data access and management software tiers

The study will hopefully yield more insights into the concrete questions of what kind of data can be migrated, and how to efficiently handle on-premises and cloud data integration. We hope to identify further issues by the fact that the prototype responds to an actual and relevant requirement in the environmental department of the authors’ institute. In the natural sciences, there is an increasing demand for data integration solutions that permit to conduct interdisciplinary research. The emerging availability of open governmental data (Murray-Rust, 2008) in this area can foster this. Furthermore, mobility of researchers and long-term persistence challenges for scientific data are additional reasons for examining cloud solutions in this domain.

Given this, this prototype is an exemplary application of data integration, and will permit to scrutinize the application of cloud computing principles in this domain.

REFERENCES

Adkinson-Orellana, L., A. Rodríguez-Silva, D., J. González-Castaño, F., 2011. Sharing Secure Documents in the Cloud. CLOSER 2011 - International Conference on Cloud Computing and Services Science.

Al-Zoube, M., 2009. E-Learning on the Cloud - International Journal of Virtual and Personal Learning Environments.

Anderson, J., Bagnall, R., Smythe, M., 2011. Position Reporting Obligations - Investment Advisers

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patter-son, D., Rabkin, A., Stoica, I., Zaharia, M., 2009. Above the Clouds: A Berkeley View of Cloud Computing.

Baars, H. and Kemper, H. G., 2010. Business Intelligence in the Cloud?. PACIS 2010 Proceedings

Berbner, R., Grollius, T., Repp, N., 2005. An approach for the Management of Service-oriented Architecture (SoA) based Application Systems - Enterprise Modelling and Information Systems Architectures.

Bernstein, P., Haas, L., 2008. Information Integration in the Enterprise.

Blodget, H., 2011, Amazon's Cloud Crash Disaster Permanently Destroyed Many Customers' Data. Retrieved January 21, 2012 from http://articles.businessinsider.com/2011-04-28/tech/29958976_1_amazon-customer-customers-data-data-loss#ixzz1l21WHWng

Böhm, C., Naumann, F., Freitag, M., George, S., Höfler, N., Köppelmann, M., Lehmann, C., Mascher, A., Schmidt., T., 2010. Linking Open Government Data: What Journalists Wish They Had Known - Proceedings of the 6th International Conference on Semantic Systems.

D'Agostino, S., Ahronovitz, M., Armstrong, J., Ahmad, R., Davalbhakta, N., Gogulapati, R., Lau, E., Luster, E., A. M. Matsui, A., Mohammed, A., Moskowitz, D.,

Nolan, M., Plunkett, T., Porwal, S., Raj Radhakrishnan, A., Richet, J.L., Prasad Rimal, B., Russell, D., B. Sigler, M., Sreenivasan, K., Stratton, P., Syputa, R., Tidwell, D., Venkatraman, K., Versace, M., 2010. Moving to the Cloud. Cloud Computing Use Cases Discussion Group.

Dekkers, M., Polman, F., Velde, R. T., Vries, M. D., 2006, Measuring European Public Sector Information Resources. Part 1: Description, overview of results and conclusions. Retrieved January 20, 2012, from http://ec.europa.eu/information_society/policy/psi/docs/pdfs/mepsir/final_report.pdf.

Doelitzscher, F., Sulistio, A., Reich, C., Kuijs, H., Wolf, D., 2011. Private Cloud for Collaboration and e-Learning Services: from IaaS to SaaS. Journal Computing - Cloud Computing archive Volume 91 Issue 1.

Géczy, P., Izumi, N., Hasida, K., 2012. Cloudsourcing: Managing Cloud Adoption - Global Journal of Business Research

Halevy, A., Rajaraman, A., Ordille, J., 2006. Data Integration: The Teenage Years. Proceeding VLDB '06 Proceedings of the 32nd international conference on Very large data bases.

Hasselbring, W., 2000. Information system integration - Communications of the ACM (Vol. 43, pp. 32-38).

Hevner, A. R., Ram, S., March, S. T., 2004. Design Science in Information Systems Research - MIS Quarterly.

Khatri, V., Brown, C., 2010. Designing Data Governance - Magazine Communications of the ACM.

Kim, W., 2009. Cloud Computing: Today and Tomorrow. Journal of Object Technology (JOT).

Lenzerini, M., 2002. Data Integration: A Theoretical Perspective - Symposium on Principles of Database Systems.

Mell, P., Grance, T., 2011. The NIST Definition of Cloud Computing. Retrieved January 24, 2012, from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

Miller, P., Styles, R., Heath, T., 2008. Open Data Commons, a License for Open Data. 1st Work-shop on Linked Data on the Web.

Murray-Rust, P., 2008. Open Data in Science. Unilever Centre, Department of Chemistry, Univeristy of Cambridge.

Naumann, F. and Rolker, C. 2000. Assessment methods for information quality criteria. Proceedings of 5th International Conference on Information Quality

Spies, A., 2011. Global Data Protection: Whose Rules Govern? Retrieved January 28, 2012, from http://www.bingham.com/Media.aspx?MediaID=12931

Yahanna, N. and Gilpin, M., 2012, The Forrester Wave: Data Virtualization, Q1, 2012. Retrieved January 28, 2012, from http://www.forrester.com/rb/Research/wave%26trade%3B_data_virtualization%2C_q1_2012/q/id/60746/t/2