UNIVERSITY
OF TRENTO

DIPARTIMENTO DI INGEGNERIA E SCIENZA DELL'INFORMAZIONE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

TOWARDS KNOWLEDGE IN THE CLOUD

Davide Cerri, Emanuele Della Valle, David De Francisco Marcos,
Fausto Giunchiglia, Dalit Naor, Lyndon Nixon, Kia Teymourian,
Philipp Obermeier, Dietrich Rebholz-Schuhmann,
Reto Krummenacher, and Elena Simperl

August 2008

Technical Report # DISI-08-030

Also: published on SEMELS'08: International Workshop on Semantic Extensions to Middleware: Enabling Large Scale Knowledge Applications; Part of OTM Conferences (COOPIS), Mexico, November 2008.

Towards Knowledge in the Cloud

Davide Cerri[1], Emanuele Della Valle[1,2], David De Francisco Marcos[3], Fausto Giunchiglia[4], Dalit Naor[5], Lyndon Nixon[6], Kia Teymourian[6], Philipp Obermeier[6], Dietrich Rebholz-Schuhmann[7], Reto Krummenacher[8], and Elena Simperl[8]

[1] CEFRIEL – Politecnico of Milano, Via Fucini 2, 20133 Milano, Italy
{davide.cerri,emanuele.dellavalle}@cefriel.it
[2] Dip. di Elettronica e Informazione, Politecnico di Milano, Milano, Italy
emanuele.dellavalle@polimi.it
[3] Telefonica Investigacion y Desarollo, Valladolid, Spain davidfr@tid.es
[4] Dipartimento Ingegneria e Scienza dell'Informazione, University of Trento, Povo, Trento, Italy fausto@disi.unitn.it
[5] IBM Haifa Research Laboratory, Haifa, Israel DALIT@il.ibm.com
[6] Institute of Computer Science, Free University of Berlin, Berlin, Germany
nixon@inf.fu-berlin.de
[7] European Molecular Biology Laboratory, European Bioinformatics Institute, Heidelberg, Germany rebholz@ebi.ac.uk
[8] Semantic Technology Institute, University of Innsbruck, Austria
{reto.krummenacher,elena.simperl}@sti2.at

Abstract. Knowledge in the form of semantic data is becoming more and more ubiquitous, and the need for scalable, dynamic systems to support collaborative work with such distributed, heterogeneous knowledge arises. We extend the "data in the cloud" approach that is emerging today to "knowledge in the cloud", with support for handling semantic information, organizing and finding it efficiently and providing reasoning and quality support. Both the life sciences and emergency response fields are identified as strong potential beneficiaries of having "knowledge in the cloud".

1 Introduction

Knowledge in the form of semantic data [2] is becoming increasingly ubiquitous in the Internet, but important steps towards scalable, dynamic systems to support collaborative work with distributed, heterogeneous knowledge are still missing. Following the *data in the cloud* paradigm that is emerging today (such as Amazon S3[3]), in this paper we propose a future vision of "knowledge in the cloud". This vision is applicable to critical collaborative tasks as different as life sciences and emergency response. In the life sciences, large scale knowledge about protein functions must be optimally organized and ubiquitously available so that

[2] promoted by the W3C Semantic Web activity http://www.w3.org/2001/sw
[3] http://aws.amazon.com/s3

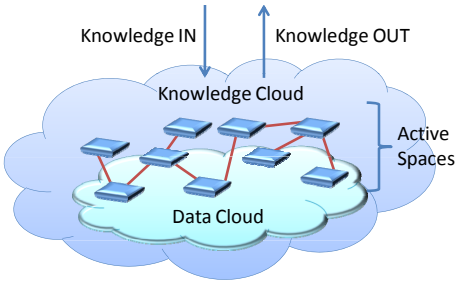

Fig. 1. The vision of knowledge in the cloud.

new inferences can be made and passed as input to collaborating analysis activities. In emergency response, deriving facts from knowledge about flooding must be available to emergency workers in a time critical manner in order that they collaborate effectively.

Data in the cloud refers to the cloud storage idea, where data is stored somewhere on the Web through abstract APIs with loose schemas and without *any* constraint of space, availability and scalability. Clients can *completely* rely on the data cloud and count on loose coupling, as access is not tied to particular access patterns dependent on the use of specific schemas. This loose coupling is similar to the one provided by Triplespace Computing [1, 2], an emerging coordination paradigm combining semantic, tuplespaces [3] and Web Service technology for the persistent publication of knowledge and the coordination of services using that knowledge.

We believe in the possibility to merge the Triplespace Computing paradigm with data in the cloud forming the "knowledge in the cloud" vision, which incorporates support for knowledge (semantic data), co-ordination (collaboration) and self-organization (internal optimisation).

"Knowledge in the cloud" is illustrated in Figure 1.

Applications are increasingly sharing their data in the cloud. To take advantage of the "knowledge in the cloud" vision, firstly the semantic knowlegde must be extracted from the underlying data. This knowledge is shared in the overlying knowledge cloud (which provides ubiquitous access) in active (i.e. with reasoning) spaces (which provide collaboration and coordination).

In the rest of the paper, we present how we believe this vision can be realized using the approaches mentioned above (Section 2). In Section 3, we provide a short description of two scenarios that would benefit from application with knowledge in the cloud. Finally in Section 4 and 5, we briefly compare our vision with other on-going work and we discussion its potential impacts.

2 Approach

In this section we introduce the approaches to be taken and applied to realise the "knowledge in the cloud" vision outlined in this paper.

2.1 Cloud Storage

Cloud Storage provides support for always available, ubiquitous access to hosted data. A virtualized entity of data is made available online and hosted on a variety of multiple virtual servers, mostly hosted by third parties, rather than being hosted on dedicated servers. Such cloud storage approaches are being increasing used in real world scenarios to make application access to data independent from a physical machine or connection.

While such solutions exist for Internet-scale, highly available data storage and management, they do not address specifically the case where that data is in fact knowledge, i.e. semantic data expressed by a formal logical model. This includes the map-reduce paradigm from Google [4] and the basic S3 cloud storage service of Amazon. Although it provides a highly available storage service that is cheap to maintain, it mainly targets large data objects that are not frequently modified and is not context aware.

To support "knowledge in the cloud", a special cloud based storage service is required which is tailored for handling semantic data, which includes allowing execution of reasoning over the data and can support distributed query processing.

2.2 Self-organizing Systems

Self-organizing systems seek to implement a decentralised, autonomous organization of data within a widely distributed system. This is able to provide stability, dynamicism and scalability. We consider these important requirements on "knowledge in the cloud", where large amounts of knowledge must be always-available in the cloud.

Swarm intelligence can be an innovative solution for "knowledge in the cloud". Swarm individuals are the active entities that are able to observe their neighborhood, to move in the environment, and to change the state of the environment in which they are located. For example, they might move between nodes in a distributed system, picking up and dropping data in their path, performing calculations or executing new processes. Despite the lack of central control, such systems demonstrate emergent global behaviour patterns.

SwarmLinda [5] is to our knowledge the only system which combines tuplespace computing and self-organization principles. SwarmLinda uses the ant colony pattern, to cluster tuples according to their content, and efficiently answer queries through following scents left by tuple-carrying ants. Ants can also load-balance tuples on nodes. The principles of SwarmLinda can be used to optimize tuple distribution and retrieval in large-scale space-based systems. First implementation results [6] demonstrate the expected clustering of data in a decentralised manner, leading to greater efficiency, dynamicism and scalability in the system.

Knowledge in the cloud must go beyond the state of the art, extending swarming to support the distribution and retrieval of semantic data, creating a new research field of semantic swarms. Now notions of fitness, relatedness, similarity

etc. must be inferenced from semantic data. By that, we aim to achieve intelligent swarm intelligence. This fundamental shift will result in new and powerful decentralized and scalable algorithms for applying swarm computing to triplespaces.

2.3 Distributed Querying and Reasoning

Distribution of a system is a necessity for achieving Internet scalability, hence distributed query processing is a requirement on "knowledge in the cloud". Research in distributed semantic query processing started recently and predominantly leans on optimization techniques for general database query languages.

There are known query processing optimization concepts in distributed databases as well as rule-based rewriting techniques for queries which deliver optimization on a logical level independent from the physical implementation. Current work in distributed semantic querying includes DARQ [7] and SemWIQ [8].

There has been no consideration of the potential of swarm-based approaches as a divide and conquer approach to query optimisation in distributed semantic systems, as we propose here for "knowledge in the cloud". The aim here is to move reasoning out of the reasoner on top of the data store, and into the cloud of self-organizing semantic data, exploring new forms of distributed reasoning where inference processes co-operating in the cloud can optimize the reasoning task for large scale knowledge clouds.

Reasoning individuals as part of the swarm can have a rather limited set of reasoning capabilities compared to a full blown theorem proofer - thus it is called micro-reasoning. We divide reasoning tasks across these individuals and apply swarm algorithms to them, such that the collective reasoning in the swarm provides inference on data expressed by common semantic models (subsets of Description Logics and Logic Programming) and therefore permits the solution of complex problems. Given the reduction of the required schema information for each individual, and the distribution of this schema information in any case, we expect that such micro-reasoners can provide more efficient and scalable reasoning services for well defined reasoning sub-tasks.

A further extension of reasoning is to support "active spaces", which allow rules to be stored in tuples besides RDF and OWL semantic information, triggering automatic generation of knowledge beyond the capabilities of the RDF and OWL languages and supporting notification to clients of new knowledge states which have been inferred from the existing knowledge.

2.4 Trust and Data Quality

The "knowledge in the cloud" can be distributed with a Peer-to-Peer (P2P) architecture. P2P brings advantages for data and knowledge management systems such as efficient routing, fault tolerance, and low cost of ownership. At the same time, it opens up new research challenges such as those related to the assessment of the quality of data. In critical collaboration scenarios as foreseen in this paper, the quality of the semantic data is vital for correct inferences from the knowledge.

The notion of data quality is well understood in the context of centralized information systems. Quality assessment in these approaches relies on the assumptions that data can be accessed and evaluated centrally, and that any piece of data in the system can be retrieved upon a user request. These data quality metrics cannot be directly applied in the P2P setting due to its decentralized, dynamic and subjective nature. In fact, in a P2P system, query results are often incomplete. Furthermore, it is hard to evaluate the correctness of a query result due to the fact that each peer maintains its subjective view on the data. Thus, the same data can be considered as correct by one peer and as incorrect by another. See [9]for a detailed discussion.

Thus, the standard data quality metrics have to be reconsidered in order to reflect the peculiar characteristics of P2P systems. However, very little has been done so far in this direction. One such case is the work done as part of the project OpenKnowledge (EU IST-27253) [10]. Here, data quality is measured by combining trust values and the result of a matching process [11–13] that aligns interaction models of different peers [14].

In "knowledge in the cloud", we must go beyond the state-of-the-art because we need to develop, implement, and evaluate a methodology for assessing the quality of answers within the knowledge cloud, a knowledge-centric self-organizing P2P system. Many spaces in the system, owned by different and unrelated entities, may be suitable for the fulfilment of a request. In this case, the requester must decide which space(s) to choose; in doing this, different information can be taken into account. A space in the system, seen as a resource, can include in its description also a quality value, which is self-declared: the owner of that resource autonomously states the quality of what he provides. A requester, after having used a resource, can evaluate the quality of that resource, and in case also give a feedback that states his opinion about the quality of that resource.

To the best of our knowledge, this methodology would be the first attempt to provide a basis for qualitative and quantitative evaluation of query answers within this kind of P2P system.

3 Application Scenarios

We identify two application scenarios that can benefit from Knowledge in the Cloud technology, and can provide valuable use cases: life science and emergency response.

3.1 Life Science

The life science community already connects a large number of scientists sharing data sets and collects results of analyses in large scale collaborative tasks. However, as these data sets become increasingly semantic (enabling new forms of inference of results) and forms of collaboration more complex, the large scale data infrastructure must support expressive knowledge models which can scale

and co-ordinate large numbers of interdependent processes over the infrastructure. We foresee "knowledge in the cloud" as fulfilling this requirement.

The proposed use case would deliver facts from the scientific literature and from the bioinformatics scientific data resources into the "knowledge in the cloud" infrastructure. Appropriate fact databases (e.g., UniProtKb [15]) and associated ontologies are both available, i.e. their content can be expressed as semantic data and thus can be fed into the knowledge cloud to make data ubiquitously available to the community, including the automatically inferred new knowledge.

In general terms, two basic research goals form the core of ongoing research in the life science community: (1) the identification of novel biological principles that explain the functioning of cells and organisms (biomedical science), and (2) the identification of agents (e.g., chemical entities, modifications to genes) that can be used to improve the functioning of cells or organisms, in particular under the condition that they malfunction. "Knowledge in the cloud" provides the required infrastructure to carry out this research.

3.2 Emergency Response

Emergency monitoring and coordination activities usually involve a range of different organizations and teams at various administrative levels with their own systems and services. In a real emergency situation, these teams need to maximally coordinate their efforts in order to work together on time-critical tasks. However, because these teams come from different organizations, they generally have incomplete or even, contradictory knowledge of the actual emergency situation. Therefore, the coordination of numerous heterogeneous actors, policies, procedures, data standards and systems results in problems in collaborative work with respect to data and knowledge analysis, information delivery and resource management, all of which are critical elements of emergency response management. Hence Knowledge in the Cloud, by being semantically rich, co-ordinated and scalable, can be very valuable in supporting such scenarios.

Emergency situations involve multiple user profiles, where each of them performs a different task. The emergency situation in our target use case, flooding, (see [14] for a decription of the scenario implementation in the OpenKnowledge project) involves about 10 different user profiles such as emergency coordinators, firefighter coordinators, police officer coordinators, medical staff, bus/ambulance drivers, and others. The platform to be developed is intended to find practical applications in municipal emergency services, where it can be used for: (i) finding points of possible failures and bottlenecks in emergency activities through simulating emergency situations; (ii) training of emergency personnel through running a simulator, thus improving their level of preparedness in real emergency situations; and, (iii) for supporting decision making procedures at runtime in a real emergency situation, in which taking an important decision quickly is an indispensable asset.

4 Related Works

The paradigm of data or computing in the cloud as an alternative to fully integrated approaches are becoming more and more prominent. Companies like Amazon with its Simple Storage Service S3[4] and the Elastic Compute Cloud EC2[5] or GigaSpaces Technologies Ltd. with its space-based computing platform XAP [16] explore the concept of cloud computing for the realization of scalable and fault-tolerant applications. Recently GigaSpaces released a cloud application server that enables their platform on top of the EC2 framework. The two companies argue that this combination enables an on-demand model of cloud computing with a usage-based pricing model. Moreover, it provides cloud portability, on-demand scalability of the cloud, self-healing, and high-performance. In contrast to our proposal however, neither Amazon nor GigaSpace yet exploit the benefits of semantic technologies.

Table 1. Related research efforts.

Project	Key Area	Relevant Technology	Application
TripCom	Triplespace Computing	Triplespaces	Basis for active spaces
Open Knowledge	Coordination semantics and models in P2P networks	Interaction models, matching of lightweight ontologies, trust and service composition via reputation	Basis for trust and data quality
LarKC	Massive distributed reasoning	Reasoning algorithms that relax completeness and correctness requirements	Incorporation in active spaces and distributed querying
Reservoir	Cloud computing	Cloud storage	Basis for cloud platform

Loosely coupled solutions to storage integration were also recognized as important by the database community. [17] propose an integration framework that does not rely on a priori semantic integration, but rather on the co-existence of data sources. Such a system – called dataspaces – delivers a layer of core functionalities on top of the actual data providers that only virtually exposes a data integration platform to applications and users. This allows applications to focus on their own functionality rather than on the challenges of data integrity and efficiency of integration. Furthermore, by keeping the individual data sources on distinctive machines conserves the advantages of distributed systems: no centralized server (thus avoiding bottlenecks with respect to performance and data access), robustness, and scalability of data volumes and the number of users.

[4] http://aws.amazon.com/s3
[5] http://aws.amazon.com/ec2

While these projects from industry and academia show again the relevance of enhanced cloud computing solutions, our vision adds another dimension by integrating semantic technologies and further results from various ongoing and past European research efforts. In Table 1, we show some selected research projects related to our 'Knowledge in the Cloud' vision. First and foremost, the TripCom (www.tripcom.org) project lays the foundations with its Triplespace platform. OpenKnowledge (www.openk.org) is a project that provides similar interaction machinery as required for 'Knowledge in the Cloud' based on a quite different approach. The project LarKC (www.larkc.eu) deals with massive reasoning over distributed knowledge in the Semantic Web, especially via relaxation of traditional inference requirements on completeness and correctness. The project Reservoir investigates cloud storage, and its results could serve as basis for our data cloud.

5 Conclusions

The realization of "knowledge in the cloud" would provide valuable benefits to several scientific and industrial communities. In this paper, we described briefly the vision of "knowledge in the cloud", outlined the approaches needed to implement it, and introduced two scenarios in which it enables the necessary collaboration of large scale and distributed knowledge. Now we consider its potential wider impact.

From the scientific point of view, the "knowledge in the cloud" vision can apply self-organization algorithms, based on swarm intelligence principles, as well as Semantic Web standards to the existing cloud storage state of the art. The semantic tuplespace community can benefit from the very scalable and ubiquitous storage infrastructure of cloud storage systems. Semantic repositories can also take advantage of the distributed reasoning capabilities envisioned. The aforementioned synergies can burst into new research lines that could be worth for these communities to explore.

From the industry impact point of view, the "knowledge in the cloud" vision goes beyond and completes currently emerging trends in the enterprise, such as cloud computing (placing data in the cloud), collaboration (wikis, blogs) and knowledge management. These trends are the main pillars on which current technology aggregators develop and promote solutions to emerging business models, which demand transparent, scalable, reliable and flexible knowledge infrastructures. The impact of such an infrastructure would also be beneficial for the final users, who could enjoy greater service ubiquity, with an eased and improved interaction with systems due to the transparent usage of knowledge instead of data.

This paper has presented the research directions for the realization of "knowledge in the cloud" as well as outlined two concrete illustrations of its potential value in concrete applications. We will continue to promote the vision and push the research areas mentioned here, towards a realization of "knowledge

in the cloud" as the natural next progression in and integration of semantic, co-ordination and cloud computing systems.

Acknowledgments

Giona McNeill, Paolo Besana, Dave Robertson, Maurizio Marchese, Juan Pane, Lorenzo Vaccari, Veronica Rizzi, Pavek Shvaiko from the Open Knowledge team are thanked for their valuable contribution. Ilya Zaijrayeu is also thanked for his feedback on these ideas.

References

1. Fensel, D.: Triple-Space Computing: Semantic Web Services Based on Persistent Publication of Information. In Aagesen, F.A., Anutariya, C., Wuwongse, V., eds.: Proc. of the IFIP Int'l Conf. on Intelligence in Communication Systems. Volume 3283 of Lecture Notes in Computer Science., Springer-Verlag (2004) 43–53
2. Simperl, E., Krummenacher, R., Nixon, L.: A Coordination Model for Triplespace Computing. In: 9th Int'l Conference on Coordination Models and Languages. (2007)
3. Gelernter, D.: Generative communication in linda. ACM Trans. Program. Lang. Syst. **7**(1) (1985) 80–112
4. Lämmel, R.: Google's mapreduce programming model – revisited. Science of Computer Programming **70**(1) (2008) 1–30
5. Charles, A., Menezes, R., Tolksdorf, R.: On the implementation of swarmlinda. In: ACM-SE 42: Proceedings of the 42nd annual Southeast regional conference, New York, NY, USA, ACM (2004) 297–298
6. Graff, D.: Implementation and Evaluation of a SwarmLinda System (Technical Report TR-B-08-06). Technical report, Free University of Berlin (2008)
7. Quilitz, B., Leser, U.: Querying distributed RDF data sources with SPARQL. In Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M., eds.: The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008, Proceedings. Volume 5021 of Lecture Notes in Computer Science., Springer (2008) 524–538
8. Langegger, A., Wöß, W., Blöchl, M.: A semantic web middleware for virtual data integration on the web. In Hauswirth, M., Koubarakis, M., Bechhofer, S., eds.: Proceedings of the 5th European Semantic Web Conference. LNCS, Berlin, Heidelberg, Springer Verlag (2008)
9. Giunchiglia, F., Zaihrayeu, I.: Making peer databases interact - a vision for an architecture supporting data coordination. In: CIA '02: Proceedings of the 6th International Workshop on Cooperative Information Agents VI, London, UK, Springer-Verlag (2002) 18–35
10. Robertson, D., Barker, A., Besana, P., Bundy, A., Chen-burger, Y.H., Giunchiglia, F., Harmelen, F.V., Hassan, F., Kotoulas, S., Lambert, D., Li, G., Mcginnis, J., Osman, F.M.N., Sierra, C., Walton, C.: Worldwide intelligent systems. In: In Advances in Web Semantics, IOS Press (1993)
11. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic matching: Algorithms and implementation. Journal on Data Semantics **1** (2007) 2007

12. Giunchiglia, F., Yatskevich, M., Giunchiglia, E.: Efficient semantic matching. In: ESWC. (2005) 272–289
13. Giunchiglia, F., Giunchiglia, F., Yatskevich, M., Yatskevich, M.: Element level semantic matching. In: In Proceedings of Meaning Coordination and Negotiation workshop at ISWC. (2004)
14. Fausto Giunchiglia, Fiona McNeill, M.Y.J.P.P.B.P.S.: Approximate structure preserving semantic matching. In: Proceedings of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2008). (2008)
15. UniProt-Consortium: Uniprot: the universal protein knowledgebase. Nucleic Acids Research **32** (2004) 115–119
16. Shalom, N.: The Scalability Revolution: From Dead End to Open Road - An SBA Concept Paper. GigaSpaces Technologie (2007)
17. Franklin, M., Halevy, A., Maier, D.: From Databases to Dataspaces: a New Abstraction for Information Management. ACM SIGMOD Record **34**(4) (2005) 27–33