

Analysis of Web-Scale Cloud Services



Talal H. Noor • *Taibah University*

Quan Z. Sheng • *University of Adelaide*

Anne H.H. Ngu • *Texas State University*

Schahram Dustdar • *Vienna University of Technology*

Cloud services have unique characteristics, including dynamic and diverse service offerings at different levels, few standardized description languages, and varied deployment platforms. Searching such services is thus challenging. The authors' cloud service crawler engine collects metadata about 5,883 cloud services over the Web after parsing more than half a million possible links. An extensive statistical analysis on this data gives an overall view of cloud service provisioning's current status.

In the past few years, cloud computing has gained considerable momentum as a new computing paradigm for provisioning diverse services. With cloud computing, large-scale, distributed workflow applications can aggregate services and scalable computing resources on demand with practically no capital investment and modest operating costs.¹⁻⁵ Despite a considerable amount of research on addressing various cloud computing challenges, cloud services discovery remains an untouched area.^{2,4}

Indeed, in the context of cloud computing, we must revisit service discovery challenges for several reasons (see the sidebar for more research on this topic). First, cloud services are offered at different levels. Currently, at least three different service levels are available – software as a service (SaaS), platform as a service (PaaS), and infrastructure as a service (IaaS). Second, the lack of standards for describing and publishing cloud services makes discovering them even harder. Unlike Web services, which use standard languages such as the Web Services Description Language (WSDL) to expose their interfaces and UDDI to publish their services to registries, the majority of publicly available cloud services aren't based on description standards,² making cloud service discovery

problematic. For example, some publicly available cloud services (such as Dropbox) don't mention "cloud" at all, whereas some businesses that have nothing to do with cloud computing (such as cloud9carwash; www.cloud9carwash.com) might use "cloud" in their names or service descriptions.

Several interesting questions center on cloud services discovery:

- How do we identify whether a service on the Web is a cloud service?
- How many cloud services are currently available on the Web, and who provides them (that is, are cloud services provided only by major vendors such as Microsoft, IBM, Amazon, Google, and so on)?
- What kind of cloud service providers are on the Web?
- From which part of the world are cloud services provisioned?
- To what extent do established service-oriented computing (SOC) standards contribute to cloud computing?
- To what extent do consumers trust cloud services?
- Is there any publicly available cloud service dataset for use in cloud computing research?

Related Work in Cloud Services Discovery

Service discovery is a fundamental approach in several research areas, including ubiquitous computing, mobile ad hoc networks, peer-to-peer (P2P), and service-oriented computing.^{1–3} However, with the advent of the cloud, we must reconsider challenges in this area because solutions for effective cloud service discovery are limited.^{1,4}

Some researchers propose ontology techniques for cloud services discovery. One study proposes a cloud service discovery system (CSDS) that exploits ontology techniques to find cloud services that are closer to consumers' requirements.⁴ Here, agents perform reasoning methods such as similarity, equivalent, and numerical reasoning. Unfortunately, this work is only validated in a small, simulated environment. We conducted our work across the entire Web to discover real cloud services. In addition, our cloud services ontology design follows the US National Institute of Standards and Technology (NIST) cloud computing standard, which helps in filtering out noisy data and increasing discovery results' accuracy.

Other researchers propose using distributed hash tables (DHTs) for better discovery and load balancing of cloud services. One study presents the concept of a cloud peer that extends DHT overlay to support indexing and matching of multidimensional range queries for service discovery.⁵ This approach is validated on a public cloud computing platform (Amazon EC2). The authors' work focuses on a closed environment. In contrast, we focus on discovering cloud services on an open environment (that is, the Web) to let any users or applications search cloud services that suit their needs.

Discovering Web services has been an active research area with some good results. One work collects Web Services Description Language (WSDL) documents by crawling UDDI business registries (UBRs) as well as search engines such as Google, Yahoo, and Baidu.² The authors present some detailed statistical information on Web services, such as active versus inactive Web services

and object size distribution. Another study collects Web services data through Google API and presents some interesting statistical information related to Web services' operation, size, word distribution, and function diversity.⁶ Most recent are findings on the current status of RESTful Web services.⁷ The authors use 17 different RESTful service design criteria (for example, availability of formal description) to analyze the top 20 RESTful services listed on the ProgrammableWeb (www.programmableweb.com). Unlike previous work that discovers Web services by simply collecting interface documents (such as WSDL files) and searching UBRs, discovering cloud services presents more challenges, such as the lack of standardized description languages for cloud services, which need full consideration.

References

1. Y. Wei and M.B. Blake, "Service-Oriented Computing and Cloud Computing: Challenges and Opportunities," *IEEE Internet Computing*, vol. 14, no. 6, 2010, pp. 72–75.
2. E. Al-Masri and Q. Mahmoud, "Investigating Web Services on the World Wide Web," *Proc. 17th Int'l Conf. World Wide Web*, 2008, pp. 795–804.
3. E. Meshkova et al., "A Survey on Resource Discovery Mechanisms, Peer-to-Peer, and Service Discovery Frameworks," *Computer Networks*, vol. 52, no. 11, 2008, pp. 2097–2128.
4. J. Kang and K.M. Sim, "Towards Agents and Ontology for Cloud Service Discovery," *Proc. 2011 Int'l Conf. Cyber-Enabled Distributed Computing and Knowledge Discovery*, 2011, pp. 483–490.
5. R. Ranjan et al., "Peer-to-Peer Cloud Provisioning: Service Discovery and Load-Balancing," *Cloud Computing: Principles, Systems, and Applications*, Springer, 2010, pp. 195–217.
6. Y. Li et al., "An Exploratory Study of Web Services on the Internet," *Proc. IEEE Int'l Conf. Web Services*, 2007, pp. 380–387.
7. D. Renzel et al., "Today's Top RESTful Services and Why They Are Not RESTful," *Proc. Web Information Systems Eng.*, LNCS 7651, 2012, pp. 354–367.

Here, we describe our design of a cloud services crawler engine (CSCE) and report our statistical analysis on 5,883 real cloud services collected from the Web.

Cloud Service Crawler Engine

Our CSCE crawls search engines and collects cloud service information available on the Web. Figure 1a shows the CSCE's system architecture, which consists of six layers.

The *cloud service providers* layer (top right in Figure 1a) consists of

different cloud service providers who publicly provision and advertise their services on the Web. These cloud services are accessible through Web portals and indexed on search engines such as Google, Yahoo, and Baidu. Some websites, such as Cloud Hosting Reviews (<http://cloudhostingreview.com.au>) and Cloud Storage Service Reviews (<http://online-storage-service-review.toptenreviews.com>) let users provide feedback. The potential set of cloud service providers that the various search engines index form the initial input to the crawler.

The *cloud services ontology* layer maintains the cloud services ontology (CSO), which contains a set of concepts and relationships that let the crawler automatically discover, validate, and categorize cloud services. This layer maintains the ontology via the *ontology updater* module.

The *cloud services seeds collection* layer collects possible cloud service seeds (that is, their URLs). The *seed collector* module considers several possible resources in search engines, such as indexed webpages, WSDL and Web Application Description

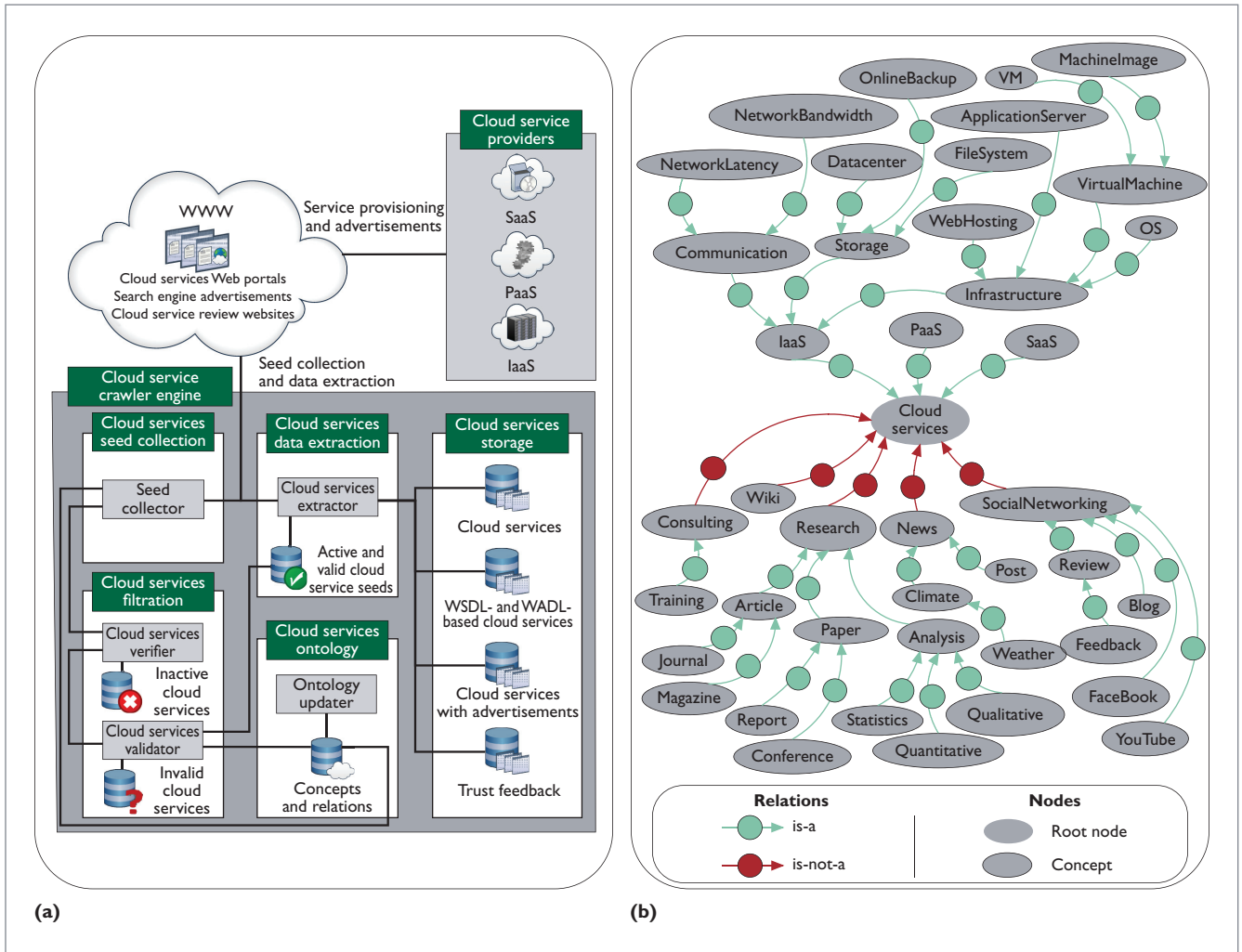


Figure 1. Cloud service crawler engine (CSCE). (a) The system architecture consists of six layers. (b) The cloud services ontology (shown in part) lets the CSCE collect possible cloud service seeds and filter out invalid ones.

Language documents (WADL is the REST equivalent of WSDL used to describe RESTful Web services), and advertisements. To collect data, the seed collector uses some of the concepts in the first few levels of the cloud services ontology as keywords (cloud services, IaaS, PaaS, SaaS, and so on), then sends the collected seeds to the *cloud services filtration* layer for validation.

The cloud services filtration layer filters the seeds collected from the seed collector. The *cloud services verifier* first determines whether a cloud service's seed is active or inactive. Inactive seeds are kept in the *inactive cloud services database* for further

checking (some inactive seeds might just be temporarily unavailable), and the error codes are also captured. Active seeds are passed to the *cloud services validator*, which validates them using concepts from the cloud services ontology. For example, if the seed's webpage contains concepts that are related to cloud services, such as IaaS, storage, and infrastructure, then the seed is considered valid. However, if the seed's webpage contains other concepts, such as news, article, paper, or weather, then the seed is invalid because the collected seed could be a news website that publishes articles about cloud services. Invalid seeds are kept in the *invalid cloud service*

database, whereas valid seeds are categorized (into IaaS, PaaS, or SaaS) before passing to the next layer.

The *cloud services data extraction* layer extracts information for active and valid cloud services (for example, cloud service ID, URL, and description). The data is stored in the corresponding databases in the *cloud services storage layer* for further statistical analysis.

Cloud Services Ontology

The CSO provides the crawler engine with meta-information and describes cloud services' common data semantics, which is critical in the sense that cloud services might not necessarily

Table 1. Breakdown of cloud services collection results.

Cloud services collection	Start page	WSDL/WADL	Ads	Total
Links parsed	617,285	1,552	637	619,474
Possible seeds	34,348	616	637	35,601
Inactive	366	57	0	423
Active	34,619	559	637	35,815
Invalid	28,736	453	0	29,189
Valid	5,883*	106	637	5,883

*Cloud services identified from the Web Services Description Language (WSDL)/Web Application Description Language (WADL) and advertisements are also included in the results.

Table 2. Error codes for inactive cloud services.

Error Code	Description	Percentage
101	The connection was reset	13.66
105	Unable to resolve the server's DNS address	1.64
107	Secure Sockets Layer protocol error	0.27
118	The operation timed out	0.27
324	The server closed the connection without sending any data	0.27
330	Content decoding failed	0.27
400	Bad request	0.82
403	Access denied	3.83
404	The requested URL / was not found on this server	10.11
500	Server error	1.37
503	The service is unavailable	0.27
504	Page not found	0.27
1005	URL does not exist	66.95
Total	—	100

use identity words (cloud, infrastructure, platform, software, and so on) in their names and descriptions. When developing the CSO, we considered the common concepts that appear in the cloud computing standard from the US National Institute of Standards and Technology (NIST; <http://csrc.nist.gov/publications/drafts/800-146/Draft-NIST-SP800-146.pdf>).

Our CSO contains a set of concepts and relationships between concepts that lets the CSCE automatically discover, validate, and categorize cloud services on the Web. We developed the CSO based on the Protégé Ontology Editor and Knowledge Acquisition System (<http://protege.stanford.edu>), which we used to construct the ontology and reason over the concepts. These concepts let the CSCE collect possible cloud service seeds and filter out invalid ones. The CSO defines two different relations: is-a and is-not-a. For instance, the seed collector uses the concepts that are associated with is-a relations (the top part of Figure 1b) to collect possible cloud service seeds from search engines. On the other hand, the cloud services validator uses concepts that are associated with is-not-a relations (the bottom part of Figure 1b) for cloud services validation. Finally, the cloud services validator uses the concepts that are associated with is-a relations to categorize a valid

cloud service as IaaS, PaaS, SaaS, or a combination of these models.

Statistical Analysis and Results

We present a comprehensive, statistical analysis of the collected data on cloud services, from several different aspects. These results also provide some insight into the questions we presented in the introduction.

Cloud Services Identification

To optimize the crawling performance, we used three different instances of the CSCE (each instance collects the data using multiple threads) to run simultaneously from three different machines. At an early stage, we configured the crawler to crawl up to five levels deep in a potential cloud service's website. However, we discontinued this because it's time consuming, and no significant difference exists in the crawling results. Therefore, we configured the crawler to crawl the first level of a potential cloud service's website, where the service description is usually found. Table 1 breaks down the cloud services collection and verification results. During collection, a significant portion of noisy data is present. After parsing 619,474 links, the crawler found 29,189 invalid seeds from 35,601 possible seeds for cloud services (more than 80 percent). This is largely attributed to the fact that we lack standards for describing and publishing cloud services. Therefore, an urgent need exists for standardization on cloud services, such as interfacing and discovery.

Note that the total number of inactive cloud services is significantly low (only 423, or roughly 0.1 percent of the total possible seeds). Search engines regularly check outdated links and exclude them from their indexes. For those inactive cloud services, our crawler also captured the error codes according to the RFC 2616 status code definitions from the W3C (www.w3.org/Protocols/rfc2616/rfc2616-sec10.html), as Table 2 shows.

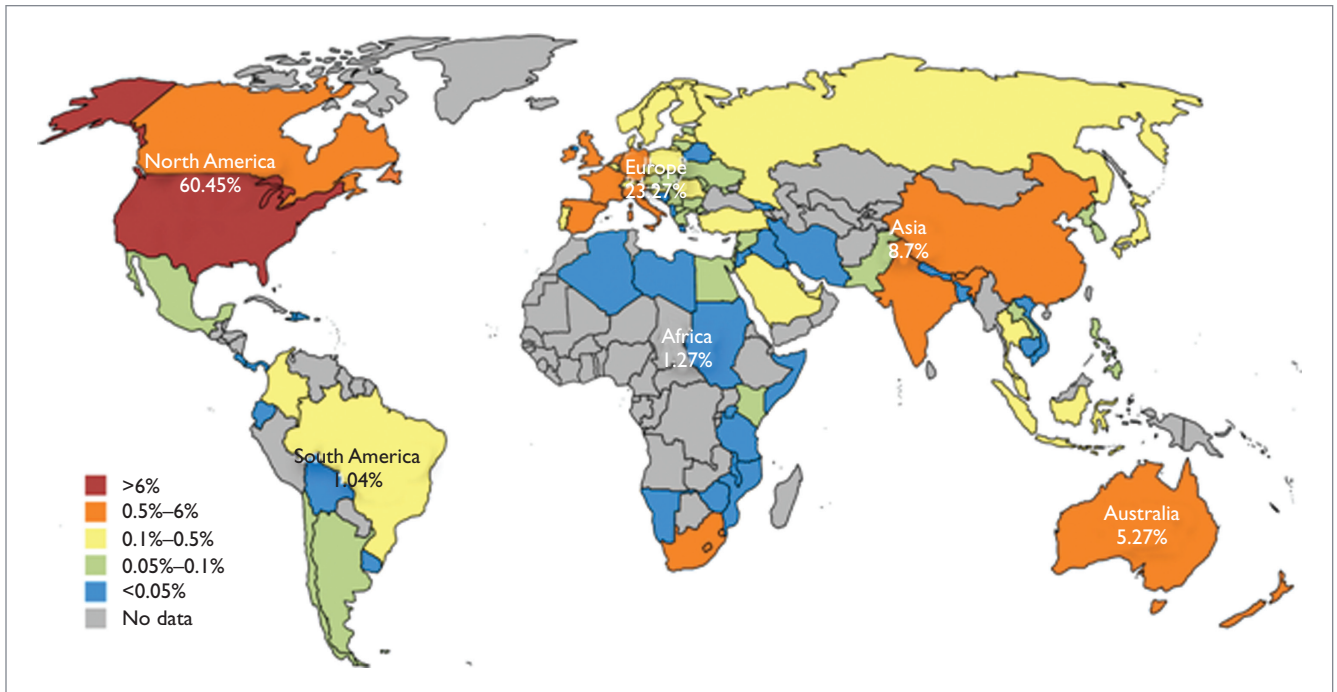


Figure 2. Cloud services locations. This gives us a holistic view of cloud computing trends worldwide.

In this table, we can see that the highest percentage (66.95 percent) goes to error code 1005 (that is, the URL doesn't exist), which means that the majority of inactive cloud services are discontinued.

Locations and Languages

One of our studies about cloud services, and cloud computing in general, deals with its geographical status (that is, from which part of the world cloud services are provisioned). We extracted the country domain from each URL of the collected cloud services. When the country domain wasn't present, we exploited address lookup tools such as whois (<http://ipduh.com/ipv6/whois/> or www.sixxs.net/tools/whois/) to determine the URL's location, which essentially traces back to the geographical location of the hosting datacenter and helps us determine the cloud service's country information. For presentation purposes, we group countries into different regions for a holistic view of cloud computing trends and depict the information on a world map (Figure 2). We present details about particular

countries in a specific color, according to the percentage range of the cloud services that country provisions.

From Figure 2, we note that the North American region is the biggest provider for cloud services, with 60.45 percent. This is followed by Europe (23.27 percent). Asia provisions about 8.7 percent of the cloud services (about 1 percent from the Middle East), and 5.27 percent are from Australia. The remaining 2.31 percent of the cloud services are provisioned from other regions, including South America and Africa.

We also conducted some statistics on the languages used for the collected cloud services. For this task, we leveraged online tools – What Language Is This (<http://whatlanguageisthis.com>) and an open source system called Language Detection Library for Java (<http://code.google.com/p/language-detection/>). Figure 3a shows the statistical information of the languages that are used in the cloud services. From the figure, it is clear that most cloud service providers use English (85.33 percent). This is consistent with

the fact that a large portion of cloud services are provided by countries in North America, Australia, and Europe, and most of them are English speaking. We can see from other languages used in cloud services – such as Chinese, French, German, and Spanish – that cloud computing is achieving broad adoption. Noticeably, 4.30 percent of cloud services are in an Arabic language.

Cloud Service Provider Categorization

Cloud services are widely categorized as IaaS, PaaS, or SaaS, provisioned by different cloud service providers. Determining the percentages of different kinds of service providers would be interesting. As described, after our CSCE finishes validating cloud service seeds, it categorizes the cloud services into IaaS, PaaS, or SaaS by reasoning over the relations between the concepts in the cloud services ontology.

Figure 3b depicts the categorization results, with providers categorized into six different categories: IaaS, PaaS, SaaS, IaaS+PaaS, IaaS+SaaS,

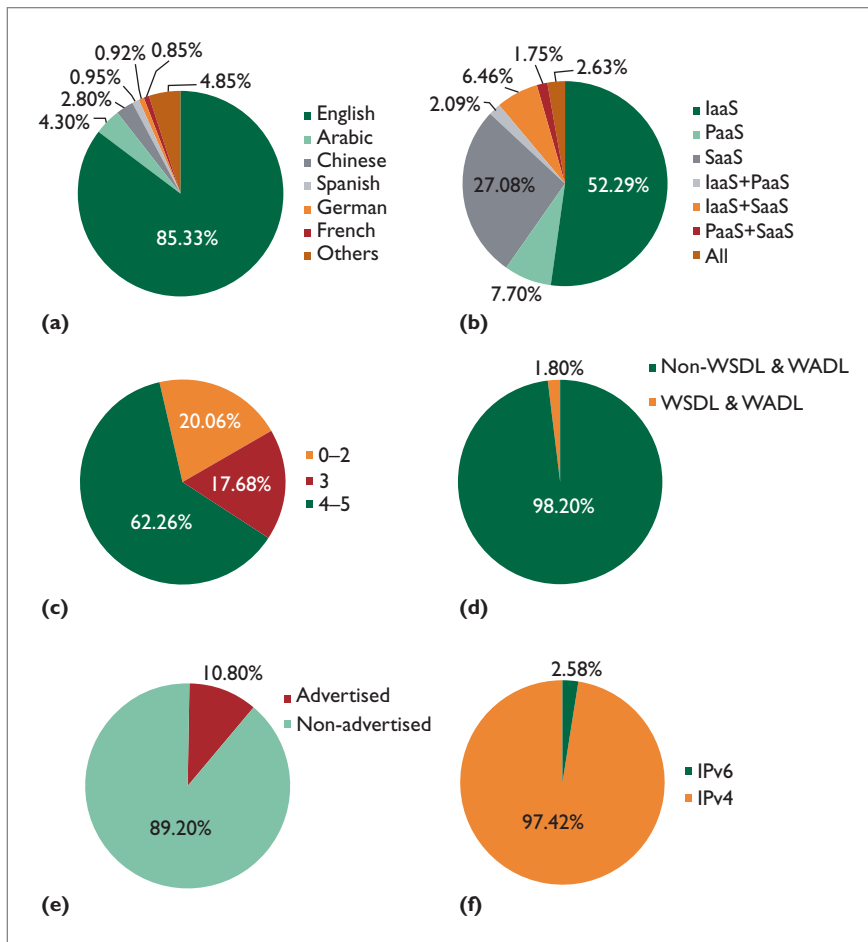


Figure 3. Statistical results and analysis. We looked at (a) languages used in cloud services; (b) cloud service provider categorization; (c) cloud service consumer trust feedback; (d) cloud services in the Web Service Description Language (WSDL) or Web Application Description Language (WADL); (e) cloud services advertised on search engines; and (f) cloud services' IPs.

PaaS+SaaS, and "all." Note that when a cloud service provider is categorized as IaaS+PaaS, this provider offers both IaaS and PaaS services. From the figure, we can see a fair degree of variation among providers. In particular, more than half (52.29 percent) focus on providing IaaS services, nearly one third (27.08 percent) on providing SaaS services, and 7.70 percent on providing PaaS services. The remaining 12.93 percent offer more than one cloud service model. Major players such as Microsoft, Amazon, and Google belong to this part.

Cloud Services and QoS

Quality-of-service (QoS) attributes are critical in cloud service discovery. With

QoS information, we could rank collected cloud services according to consumers' requirements, and always select the best ones for users or workflow applications. Our CSCE collected cloud services' QoS data by visiting review websites that document consumers' feedback. Among QoS attributes, we're particularly interested in trust, given that it's widely considered a key challenge in cloud adoption.^{1,4,6,7}

We analyzed 10,076 feedbacks collected from 6,982 users on 113 real cloud services. Figure 3c depicts the results. Cloud service consumers gave trust feedback in numerical form, with a range between 0 and 5, where 0 and 5 mean the most negative and the most

positive, respectively. From the figure, we can observe that the majority of cloud service consumers (62.26 percent) are positive (scoring 4–5) in trusting the cloud services they used. Only 20.06 percent of cloud service consumers were negative (scoring between 0–2) in trusting cloud services, and the rest (17.68 percent) of the feedback was neutral.

Cloud Services and SOC

Service-oriented computing (SOC) and Web services are one of the most important enabling technologies for cloud computing.^{2,5,8} Thus, we wanted to investigate SOC adoption in cloud computing. We conducted some preliminary studies based on the information we collected. We first investigated how much SOC description languages such as WSDL or WADL have been used for publishing cloud services. To do this, we compared the number of cloud services that have WSDL or WADL documents (for SOAP-based or RESTful Web services, respectively).

Figure 3d depicts the result. We were surprised to discover that only a very small portion of cloud services (merely 1.80 percent) were implemented using Web service interface languages. However, our crawler might not detect cloud services that actually used SOC because not all WSDL documents are publicly accessible on the Internet.⁹ In addition, the majority of RESTful Web services provide no formal descriptions and rely on informal documentation.¹⁰ Nevertheless, the low percentage still indicates poor adoption of SOC in cloud computing.

Advertisements and IPs

We also investigated how cloud services advertise themselves so that potential customers can find them. Search engines not only index cloud services, but some providers also advertise their services via this medium. These advertisements are usually located on the top or to the right of the returned search pages. Accordingly, our CSCE collected these advertised cloud services. Figure 3e

shows that about 10.80 percent of collected cloud services use paid advertisements as a means for customers to discover them. Because advertised cloud services rely only on a short description text to introduce themselves, user queries that normally require more information (for example, functions and QoS information) can't be answered via these advertisements.

Another interesting and important aspect worth investigating is the cloud services communication (that is, what type of IPs do cloud services use?). We used an `nslookup` command to determine what type of IP cloud services are using (IPv4 or IPv6). We wrote a simple Java program to enable automatic retrieval of such IP addresses from the collected URLs. As Figure 3f shows, most cloud services (97.42 percent) use IPv4. This makes sense because IPv4 is still the most widely deployed Internet-layer protocol.

The most intriguing finding in our evaluation is that SOC isn't playing a significant role in enabling cloud computing as a technology; this is contrary to what's documented in current literature. More investigation is needed to understand why this is the case and how to enable SOC to contribute toward cloud computing so as to capitalize on previous efforts in R&D in SOC communities.

In addition, the lack of standardization in current cloud products and services makes cloud services discovery a more difficult task and a barrier for scalable and unified access to such services. An urgent need thus exists for standardization, especially in description languages, before we can fully embrace cloud computing. Fortunately, the research community is making some attempts at standardization, and has achieved some initial results. For example, in August 2012, the Distributed Management Task Force (DMTF) released the Cloud Infrastructure Management Interface

(CIMI) specification, which standardizes interactions between cloud environments to achieve interoperable cloud infrastructure management (www.dmtf.org/news/pr/2012/8/dmtf-releases-specification-simplifying-cloud-infrastructure-management).

To the best of our knowledge, ours is the first effort in discovering, collecting, and analyzing cloud services on a Web scale. The collected datasets (1.06 Gbytes of metadata), which are available at <http://cs.adelaide.edu.au/~cloudarmor/ds.html>, will bring significant benefits to the cloud service research community.

Our ongoing research includes further investigating the relationship between SOC and cloud computing by discovering more evidence on cloud services implemented using SOC technology. We also plan to extend the CSCE to perform more comprehensive QoS metrics to rank cloud services. □

Acknowledgments

Talal H. Noor's work is supported by King Abdullah's Postgraduate Scholarships, the Ministry of Higher Education: Kingdom of Saudi Arabia. Quan Z. Sheng's work is partially supported by Australian Research Council Discovery grant DP130104614 and DP140100104. We thank Jeriel Law and Abdullah Alfazi for their participation in data collection.

References


1. M. Armbrust et al., "A View of Cloud Computing," *Comm. ACM*, vol. 53, no. 4, 2010, pp. 50–58.
2. Y. Wei and M.B. Blake, "Service-Oriented Computing and Cloud Computing: Challenges and Opportunities," *IEEE Internet Computing*, vol. 14, no. 6, 2010, pp. 72–75.
3. K. Ren et al., "Security Challenges for the Public Cloud," *IEEE Internet Computing*, vol. 16, no. 1, 2012, pp. 69–73.
4. T.H. Noor et al., "Trust Management of Services in Cloud Environments: Obstacles and Solutions," *ACM Computing Surveys*, vol. 46, no. 1, 2013, article no. 12.
5. B. Satzger et al., "Winds of Change: From Vendor Lock-In to the Meta Cloud," *IEEE Internet Computing*, vol. 17, no. 1, 2013, pp. 69–73.
6. T.H. Noor and Q.Z. Sheng, "Credibility-Based Trust Management for Services in Cloud Environments," *Proc. 9th Int'l Conf. Service-Oriented Computing*, 2011, pp. 328–343.
7. K. Hwang and D. Li, "Trusted Cloud Computing with Secure Resources and Data Coloring," *IEEE Internet Computing*, vol. 14, no. 5, 2010, pp. 14–22.
8. T.H. Noor and Q.Z. Sheng, "Trust as a Service: A Framework for Trust Management in Cloud Environments," *Proc. 12th Int'l Conf. Web Information System Eng.*, 2011, pp. 314–321.
9. E. Al-Masri and Q. Mahmoud, "Investigating Web Services on the World Wide Web," *Proc. 17th Int'l Conf. World Wide Web*, 2008, pp. 795–804.
10. D. Renzel et al., "Today's Top RESTful Services and Why They Are Not RESTful," *Proc. Web Information Systems Engineering*, LNCS 7651, 2012, pp. 354–367.

Talal H. Noor is an assistant professor in the Department of Computer Science at Taibah University, Yanbu. He has a PhD in computer science from the University of Adelaide. Contact him at tnoor@taibahu.edu.sa.

Quan Z. Sheng is an associate professor and head of the Advanced Web Technologies Research Group in the School of Computer Science at the University of Adelaide. He has a PhD in computer science from the University of New South Wales. Contact him at qsheng@cs.adelaide.edu.au.

Anne H.H. Ngu is a full professor in the Department of Computer Science at Texas State University. She has a PhD in computer science from the University of Western Australia. Contact her at angu@txstate.edu.

Schahram Dustdar is a full professor of computer science and head of the Distributed Systems Group, Institute of Information Systems, at the Vienna University of Technology. He is an ACM Distinguished Scientist and IBM Faculty Award recipient. Contact him at dustdar@dsg.tuwien.ac.at.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.