Confidentiality and Performance for Cloud Databases

Author(s):
Braun-Löhrer, Lucas Victor

Publication Date:
2017

Permanent Link:

https://doi.org/10.3929/ethz-a-010866596 →

DISS. ETH NO. 24055

Confidentiality and Performance for Cloud Databases

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

LUCAS VICTOR BRAUN-LÖHRER

Master of Science ETH in Computer Science, ETH Zurich

born on June 18, 1987

citizen of Aadorf, Thurgau

accepted on the recommendation of

Prof. Dr. Donald A. Kossmann, examiner
Prof. Dr. Timothy Roscoe, co-examiner
Prof. Dr. Renée J. Miller, co-examiner
Prof. Dr. Thomas Neumann, co-examiner

2017

Abstract

The twenty-first century is often called the century of the information society. The amount of collected data world-wide is growing exponentially and has easily reached the order of several million terabytes a day. As a result, everybody is talking about “big data” nowadays, not only in the research communities, but also very prominently in enterprises, politics and the press. Thanks to popular services, like Apple iCloud, Dropbox, Amazon Cloud Drive, Google Apps or Microsoft OneDrive, cloud storage has become a (nearly) ubiquitous and widely-used facility in which a huge portion of this big data is stored and processed. This makes cloud computing one of the most important, but also most exciting technologies in this present age.

In the course of the cloud’s adoption by industry, several cloud service models have developed. Starting with Infrastructure-as-a-Service (IaaS), the most basic model, the cloud service stack was extended to Platform-as-a-Service (PaaS) and finally Software-as-a-Service (SaaS). Most recently, a new class of PaaS has evolved: Database-as-a-Service (DaaS), often also simply called *cloud databases*. Cloud databases provide developers with the ease of use of a well-known SQL (or NoSQL) interface which they can program their applications against. Moreover, they make the tempting promise that developers do not need to worry about the availability, performance or confidentiality of the database, let alone the operating system or hardware it runs on, because this is all ensured by the cloud provider.

Despite all these advantages, a lot of companies are still reluctant to move to the cloud. According to CloudTweaks, a famous cloud computing blog, the two major arguments for not adopting the cloud model are price and security concerns. Both of these arguments stem from the fact that today’s cloud databases do not fully keep the aforementioned promises. First, they do not optimize the performance of database queries and updates as much as they could, which means they use more resources than needed and hence

charge the customers too much. Second, the typical measures employed to address security are the use of SQL access control, secure connections and protocols like HTTPS and SSH. While this measures do satisfactorily protect the data against adversaries from outside, there is nothing that prevents the cloud provider herself to sneak into it. Even worse, in some countries, cloud providers can be legally forced by a court to disclose the data of their customers.

Hence, this dissertation studies the performance and confidentiality of cloud databases and proposes solutions that allow cloud providers to get closer to keeping their promises and to make potential customers more confident that moving to the cloud is the right thing to do. The first part of the dissertation argues why existing cloud databases are not properly optimized for modern workloads which typically comprise a demanding mix of transactions and real-time analytics. It studies the design space for scalable, elastic database systems capable to handle such workloads and argues why certain combinations of design decisions are more favorable than others. This part also presents two concrete system implementations and shows selected results that not only demonstrate their superiority over state-of-the-art competitors, but also their potential to enhance current cloud databases with regard to performance and ultimately costs.

The second part of this dissertation studies the usefulness of different encryption schemes for cloud databases. As existing schemes are either secure, but low-performance or the other way round, we propose a novel technique that can trade-off between confidentiality and performance and therefore achieves an excellent compromise between the two. In addition, we also show how to increase the value of a cloud database by processing data across tenants confidentially, which helps cloud providers to compensate for some of their costs and therefore offer better prices to their customers.

Zusammenfassung

Das 21. Jahrhundert wird oft auch als das Jahrhundert der Informationsgesellschaft bezeichnet. Die Menge der weltweit gesammelten Daten wächst exponentiell und hat bis zum heutigen Tag ohne Weiteres eine Grösse von mehreren Millionen Terabytes pro Tag erreicht. Als Folge davon ist der Begriff „Big Data“ heute in aller Munde, nicht nur in Forscherkreisen, sondern auch in Unternehmen, in der Politik und massgeblich in der Presse. Dank breit bekannten Diensten wie Apple iCloud, Dropbox, Amazon Cloud Drive, Google Apps oder Microsoft OneDrive, ist Cloud-Speicher zu einem (beinahe) allgegenwärtigen und viel genutzten Medium geworden, welches einen Grossteil dieser Daten speichert und zur Verarbeitung zur Verfügung stellt. Nicht zuletzt aufgrund dieser Tatsache lässt sich „Cloud Computing“ als eine der bedeutendsten und spannendsten Technologien der Gegenwart bezeichnen.

Im Zuge der Einbindung der Cloud in verschiedene Geschäftszweige, hat sich ein mehrschichtiges Cloud-Dienstleistungsmodell entwickelt. Ursprünglich als einfache Infrastrukturdienstleistung („Infrastructure as a Service“ – IaaS) konzipiert, kamen über die Zeit weitere Schichten, wie jene der Plattform („Platform as a Service“ – PaaS) und schliesslich der Anwendung („Software as a Service“ – SaaS), hinzu. Eine besonders weit verbreitete Ausprägung der Plattformschicht ist dabei die *Cloud-Datenbank* („Database as a Service“ – DaaS). Cloud-Datenbanken stellen dem Entwickler eine einfache, wohlbekannte SQL- (oder NoSQL-) Schnittstelle zur Verfügung, welche zur Programmierung von Applikationen verwendet werden kann. Des Weiteren versprechen Cloud-Datenbanken, dass Entwickler sich weder um Verfügbarkeit, noch um Leistung oder Datensicherheit in der Datenbank kümmern, geschweige denn eine Ahnung vom Betriebssystem oder der Hardware haben müssen, denn all das liegt in der Verantwortung des Cloud-Anbieters.

Trotz all dieser Vorteile, zögern viele Unternehmen immer noch damit, die Cloud zu benutzen. Gemäss CloudTweaks, einem bekannten Cloud-Computing-Blog, sind die beiden meist genannten Gründe für diese Zurückhaltung Preis- und Sicherheitsbedenken. Beide Gründe stehen in einem engen Zusammenhang mit der Tatsache, dass die heutigen Cloud-Anbieter ihre Versprechen noch nicht vollständig erfüllen. Einerseits wird noch nicht sämtliches Optimierungspotential betreffend der Leistung von Datenbankoperationen genutzt, was dazu führt, dass mehr Ressourcen verwendet werden als tatsächlich benötigt und im Endeffekt der Kunde zu viel bezahlt. Andererseits bestehen die Massnahmen bezüglich der Datensicherheit aus SQL-Zugangskontrolle, sicheren Verbindungen und Protokollen wie HTTPS oder SSH. Während diese Methoden zwar ausreichen, um die Daten vor einem externen Angreifer zu schützen, so gibt es nichts, was den Cloud-Anbieter selbst daran hindert, in den Daten herumzuschnüffeln. Noch problematischer ist in diesem Zusammenhang die Tatsache, dass es Länder gibt, in denen die Regierung die Cloud-Anbieter per Gerichtsentscheid dazu zwingen kann, die Daten ihrer Kunden preiszugeben.

Daher untersucht diese Dissertation die Leistung und die Datensicherheit von Cloud-Datenbanken und schlägt einige Lösungen vor, die es Cloud-Anbietern ermöglichen sollten, näher an das Einhalten ihrer Versprechen zu gelangen und so ihre Kunden zu überzeugen, dass die Cloud auch für sie die richtige Plattform ist. Der erste Teil der Abhandlung erklärt, wo es bei bestehenden Cloud-Anbietern noch Optimierungspotential gibt, insbesondere was das Verarbeiten von grossen, laufend hereinkommenden Datenmengen und das gleichzeitige Analysieren dieser Daten in Echtzeit betrifft. Es werden verschiedene Lösungsansätze für skalierbare, elastische Datenbanksysteme für solche Anforderungen diskutiert und es wird argumentiert, welche Kombinationen von Lösungsansätzen am sinnvollsten sind. In diesem Teil werden auch zwei konkrete Systeme vorgestellt und ausgewählte Resultate gezeigt, die nicht nur die Überlegenheit dieser Systeme gegenüber anderen Produkten auf dem neuesten Stand der Technik belegen, sondern auch aufzeigen, wie künftige Cloud-Datenbanken leistungsfähiger und somit kostengünstiger gemacht werden können.

Der zweite Teil der Abhandlung beschäftigt sich mit der Nützlichkeit verschiedener Verschlüsselungs-Algorithmen, welche für Cloud-Datenbanken in Frage kommen. Weil bestehende Algorithmen meist entweder sicher und leistungsschwach oder leistungsstark und unsicher sind, schlagen wir eine neuartige Technik vor, welche Leistung und Sicherheit gegeneinander abwägen kann und somit einen exzellenten Kompromiss zwi-

schen diesen beiden Eigenschaften erreicht. Ausserdem wird auch noch aufgezeigt, wie der Wert von Cloud-Datenbanken gesteigert werden kann, indem Daten von mehreren Kunden zusammen (in einer vertraulichen Art und Weise) analysiert werden. Dieser gesteigerte Wert hilft wiederum den Cloud-Anbietern, einen Teil ihrer Kosten zu amortisieren und damit ihren Kunden günstigere Preise anzubieten.