Sichere Dateninfrastrukturen in der Forschung

Martin Weise

February 2023

Zusammenfassung

Fortschreitende Digitalisierung führt zur Erhebung und Speicherung von Daten in jedem Lebensbereich. Viele dieser (oft sensiblen) Datensätze sind lokal bei Dateneigentümern vorhanden, jedoch unzugänglich für den einzelnen Forscher. Einsicht in Kombination mit Möglichkeiten zur Analyse dieser Datensätze kann besonders hilfreich für die Beantwortung von Forschungsfragen sein, gestaltet sich aber aufwändig um sicherzustellen dass keine Daten die Infrastruktur des Dateneigentümers verlassen können.

Kontrolle und Schutz von sensiblen Daten bei gleichzeitiger Vergabe von Zugriff an Dritte scheint nicht nur im Widerspruch zu stehen, sondern die technische Umsetzung stellt auch eine bedeutsame Herausforderung für Dateneigentümer dar. Sichere Dateninfrastrukturen die Datenbesuche in einer überwachten und kontrollierten Umgebung, die, falls ordnungsgemäß errichtet und betrieben, können hohe Sicherheitsgarantien durch technische-, juristische- und prozessgetriebene Mechanismen ermöglichen. Obwohl es weltweit viele sichere Dateninfrastrukturen gibt die in unterschiedlichen Disziplinen eingesetzt werden, können die technischen Konzepte und grundlegende Abläufe Zugriff auf sensible Daten zu erhalten mit dem Five Safes Framework[1] beschrieben werden. Dieses Bezugssystem modelliert Datenzugriffe in fünf Risikodimensionen zum Forschungsvorhaben, den Personen, den Daten, des Umfelds und den Ergebnissen die durch Zugriff auf sensible Daten entstehen.

Um die Existenz der sensiblen Daten überhaupt zu kennen, muss ein Dateneigentümer, der sensible Daten Dritten zur Verfügung stellen möchte vorerst Beschreibungen (*Metadaten*) zu jedem Datensatz in einem Katalog bekanntmachen. Diese Metadaten enthalten spezifische Informationen über den Datensatz die dabei Helfen einen geeigneten Datensatz besser zu finden, Metadaten enthalten keine sensiblen Daten.

Alles startet mit dem Forscher der eine Forschungsfrage definiert und anschließend in dem Datenkatalog nach einem Datensatz sucht, der bei der Beantwortung der Forschungsfrage die notwendige Datenevidenz liefert. Falls zur Beantwortung mehrere Datensätze notwendig sind, müssen mehrere Ansuchen gestellt werden und der Dateneigentümer prüft genau ob durch die Verlinkung

[1] T., Desai, F., Ritchie, & R., Welpton. (2016). Five Safes. Designing Data Access for Research.

ungewollte Querverbindungen möglich sind, die nicht relevant zur Forschungsfrage sind. Es ist möglich, dass eine Kommission über die Zulässigkeit der Forschungsfrage entscheiden muss. Um Zugang zu den sensiblen Daten zu erhalten, müssen Forscher einen Antrag beim Dateneigentümer stellen. Dieser umfasst neben dem Datensatz der analysiert werden soll:

1. Forschungsfragen,
2. Informationen zur Identität des Forschers,
3. Liste der Applikationen die für die Analyse notwendig sind.

Nach erfolgreicher Prüfung erhält der Forscher Zugriff auf den für die Beantwortung der Forschungsfrage relevanten Teil des Datensatzes für einen bestimmen Zeitraum. Damit die Daten allerdings nicht vom Dateneigentümer zum Forscher abfließen können (was einem Verlust der Kontrolle über den Datensatz gleichkommt), gibt es zwei etablierte Methoden.

Methode 1. Der relevante Teildatensatz wird auf einem isolierten Computer in einem physischen Sicherheitsraum zur Verfügung gestellt auf dem alle für die Analyse notwendigen Applikationen bereits installiert sind. Darunter fallen auch Textsatz-Applikationen zum Verfassen von Berichten. Bis auf verifizierte Lizenzserver dieser Applikationen sind alle Verbindungen zum Internet gesperrt und es kann kein Massenspeicher an diesen Computer angeschlossen werden, die Daten verlassen die Infrastruktur nicht. Zusätzlich wird dieser Computer überwacht und Zugang zu diesem Sicherheitsraum wird nur unter Aufsicht gestattet.

Methode 2. Der relevante Teildatensatz wird auf einer virtuellen Schreibtischoberfläche zur Verfügung gestellt die ähnlich zur ersten Methode keinerlei Verbindungen außerhalb der Infrastruktur zulässt, allerdings muss der Forscher keinen physischen Sicherheitsraum besuchen, sondern kann mittels verschlüsselten Netzwerkverbindungen und der virtuellen Schreibtischoberfläche so arbeiten, als wäre er physisch in dem Sicherheitsraum anwesend. Diese Umgebung kann durch Überwachung der Ein- und Ausgaben, Aufzeichnung des Videokanals durch den Dateneigentümer ob die durchgeführten Analysen weiterhin zur Forschungsfrage passen ständig und jederzeit geprüft und falls notwendig der Zugriff sofort entzogen werden.

Homomorphe Verschlüsselung erlaubt eine mathematische Funktion auf verschlüsselten Daten durchzuführen ohne die unveschlüsselten Daten zu kennen. Das ist möglich aufgrund der *homomorphen* Eigenschaften der verschlüsselten Daten die es erlaubt Berechnungen so auszuführen, als wären sie auf den unverschlüsselten Daten ausgeführt worden, die Ergebnisse sind gleich. Dadurch entstehen sehr hohe Sicherheitsgarantien für sensible Daten. Da Forscher, besonders Angehörige von anerkannten, wissenschaftlichen Einrichtungen grundsätzlich nicht böswillig sind, werden in sicheren Dateninfrastrukturen andere Sicherheitsmodelle verwendet die im Vergleich zur homomorphen Verschlüsselung mehr Flexibilität in der Analyse zulässt.

Ein Ansatz um die Kontrolle und Schutz von sensiblen Daten bei gleichzeitiger Vergabe von Zugriff an Dritte zu realisieren wurde Februar 2022 von einem Team der Forschungsgruppe Data Science an der TU Wien publiziert[2], die sich mit der technischen Beschreibung und prototypischen Implementierung einer quelloffenen sicheren Dateninfrastruktur beschäftigt. Der Kern des Ansatzes ist eine Luftbrücke zwischen einem zentralen Datenserver, auf dem sensible Datensets verfügbar sind und der restlichen Dateninfrastruktur. Dieser zentrale Datenserver steht in einem separaten, veschlossenen Serverschrank zu dem nur ein vertrauenswürdiger Techniker Zugang hat, wobei der Zugang zu dem Serverraum selbst von einem anderen Techniker im Vier-Augen-Prinzip gestattet wird. Infrastruktur-interne Verbindungen, um die Luftbrücke zu schließen sind nur zum Zweck des Einspielen von neuen sensiblen Daten, Kopieren eines relevanten Teildatensets auf die virtuelle Schreibtischoberfläche und zum Durchführen von kritischen Sicherheitsaktualisierungen erlaubt und auch nur für die Dauer dieser Operationen. Der Zugriff auf Daten folgt Methode 2 und erlaubt Forschern das Arbeiten mit sensiblen Daten von überall aus der Welt durch die Verwendung eines Mehrschichtigen Konzeptes durch verschlüsselten Netzwerkverbindungen, -Bildübertragungsprotokolle und virtueller Schreibtischumgebungen auf denen bereits die angeforderten Applikationen und der relevante Teildatensatz vorhanden sind. Nachdem der Forscher mit der Analyse des Datensatzes fertig ist, kann beim Dateneigentümer eine Ausfuhr von Ergebnissen (Bericht, Grafik, trainiertes Machine-Learning Modell) beantragt werden, sofern keine sensiblen Daten das System verlassen.

Kurz-CV

Martin Weise ist Projektassistent an der Technischen Universität Wien. Seine Forschung befasst sich mit sicheren Dateninfrastrukturen und Management von Forschungsdaten. Derzeit entwickelt er zusammen mit einem Team der Universität Wien ein neuartiges Repositorium für Daten in Datenbanken.

[2]M., Weise, F., Kovacevic, N., Popper & A., Rauber. (2022). OSSDIP. Open Source Secure Data Infrastructure and Processes Supporting Data Visiting. Data Science Journal, 21(1), p.4. http://doi.org/10.5334/dsj-2022-004