

Bloomfilter-Generierung

Die Bloomfilter-Generierung findet am Standort statt. Die konkrete Konfiguration für den Produktivbetrieb ist der federated Trusted Third Party (fTTP) nicht bekannt. Die fTTP kann jedoch bei der Erstellung einer Schablone unterstützen und Empfehlungen bei den verwendeten Verfahren geben. Für den E-PIX als Identitätsmanagement kann die jeweiligen Domänenkonfiguration erzeugt und bereitgestellt werden. Die Geheimnisse/Passwörter müssen durch den Standort eingetragen werden, damit diese der fTTP unbekannt bleiben. Dabei müssen alle Standorte dieselben Geheimnisse/Passwörter verwenden. Die Verfahren selbst stellen kein Geheimnis dar und dürfen bekannt sein. Außerdem können Testdaten und eine Testkonfiguration bereitgestellt werden, sodass die Standorte die jeweils verwendeten Softwarelösungen auf die korrekte Erzeugung von Bloomfiltern testen können. Dies muss bei allen Standorten auf dieselbe Weise passieren, damit die Bloomfilter miteinander abgeglichen werden können. Für die Erkennung von Dubletten bzw. Personen standortübergreifend erkennen zu können, ist die korrekte Erzeugung von Bloomfiltern essentiell. Es muss dabei nicht dieselbe Softwarelösung verwendet werden, jedoch müssen diese kompatibel sein.

Bloomfilter allgemein

Ein Bloomfilter ist ein Bit-Vektor, der initial mit Nullen besetzt ist. In Abhängigkeit des Verfahrens, werden die zu codierenden Attribute so umgerechnet, dass diese mehrere Positionen im Bloomfilter auf Eins setzen. Diese Transformation kann nicht ohne Weiteres zurückgeführt werden. Anders als bei anderen Hashing-Verfahren, können die resultierenden Bloomfilter miteinander auf Ähnlichkeit verglichen werden (andere Hashing-Verfahren führen auch bei sehr ähnlichen Eingabewerten zu völlig unterschiedlichen Hashwerten). Um die Re-Identifizierung zu erschweren können entsprechende Verfahren (z.B. Random-Hashing) verwendet werden. Zusätzlich bieten Härtungsverfahren (z.B. Balanced Bloomfilter) einen höheren Schutz, weil bestimmte Angriffsvektoren erschwert oder verhindert werden.

Folgende Publikationen geben einen Überblick zu Bloomfiltern und einer Auswahl an Härtungsverfahren:

1. Bloom BH. Space/time trade-offs in hash coding with allowable errors. Commun ACM. 1970;13(7):422–6. <http://crystal.uta.edu/~mcguigan/cse6350/papers/Bloom.pdf>
2. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inf Decis Mak. 2009 Aug 25;9:41. <https://bmcmmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-9-41>
3. Schnell R, Borgs C. Randomized Response and Balanced Bloom Filters for Privacy Preserving Record Linkage. In 2016. p. 218–24. <https://ieeexplore.ieee.org/document/7836669>

Bloomfilter-Konfiguration

Um Bloomfilter auf dieselbe Art und Weise zu erzeugen, findet eine Normalisierung der Daten statt. Dies umfasst z.B. das Ersetzen von Umlauten und das Entfernen von Leerzeichen.

Folgende Attribute werden verwendet: Vorname, Nachname, Geschlecht, Geburtsdatum

Normalisierung für das Attribut Vorname

Beispiel: „Prof. Dr. rer. nat., Dr. rer. med. Anna-Marie_?“

Normalisierung	Beispiel
----------------	----------

Dr. entfernen	„Prof. rer. nat., rer. med. Anna-Marie_?“
med. entfernen	„Prof. rer. nat., rer. Anna-Marie_?“
Dipl. entfernen	„Prof. rer. nat., rer. Anna-Marie_?“
œ durch o ersetzen	„Prof. rer. nat., rer. Anna-Marie_?“
æ durch a ersetzen	„Prof. rer. nat., rer. Anna-Marie_?“
Prof. entfernen	„ rer. nat., rer. Anna-Marie_?“
rer. Entfernen	„ nat., Anna-Marie_?“
, entfernen	„ nat. Anna-Marie_?“
- entfernen	„ nat. AnnaMarie_?“
nat. entfernen	„ AnnaMarie_?“
Ing. entfernen	„ AnnaMarie_?“
? entfernen	„ AnnaMarie_“
Alle Buchstaben durch Großbuchstaben ersetzen	„ ANNAMARIE_“
Alle Umlaute ersetzen (Ä -> AE, Ü -> UE, Ö -> OE, ß -> SS)	„ ANNAMARIE_“
Diakritika entfernen (â -> a, é -> e, ...)	„ ANNAMARIE_“
Leerzeichen am Anfang und am Ende entfernen	„ANNAMARIE_“
Zeichen die nicht Teil des folgenden Alphabets sind, werden entfernt: „ABCDEFGHIJKLMNOPQRSTUVWXYZ1234567890 “	„ANNAMARIE“

Normalisierung für das Attribut Nachname

Beispiel: „Prof. Dr. rer. med. Dušek Müller“

Normalisierung	Beispiel
Dr. entfernen	„Prof. rer. med. Dušek Müller“
med. entfernen	„Prof. rer. Dušek Müller“
Dipl. entfernen	„Prof. rer. Dušek Müller“
œ durch o ersetzen	„Prof. rer. Dušek Müller“
æ durch a ersetzen	„Prof. rer. Dušek Müller“
Prof. entfernen	„ rer. Dušek Müller“
rer. Entfernen	„ Dušek Müller“
, entfernen	„ Dušek Müller“
- entfernen	„ Dušek Müller“
nat. entfernen	„ Dušek Müller“
Ing. entfernen	„ Dušek Müller“
? entfernen	„ Dušek Müller“
Alle Buchstaben durch Großbuchstaben ersetzen	„ DUŠEK MÜLLER“
Alle Umlaute ersetzen (Ä -> AE, Ü -> UE, Ö -> OE, ß -> SS)	„ DUŠEK MUELLER“
Diakritika entfernen (â -> a, é -> e, ...)	„ DUSEK MUELLER“
Leerzeichen am Anfang und am Ende entfernen	„DUSEK MUELLER“

Zeichen die nicht Teil des folgenden Alphabets sind, werden entfernt: „ABCDEFGHIJKLMNOPQRSTUVWXYZ1234567890 “	„DUSEK MUELLER“
--	-----------------

Normalisierung für das Attribut Geburtsdatum

Beispiel: „02.03.1986“

Normalisierung	Beispiel
Format: JJJJMMTT JJJJ – Jahr MM – Monat TT – Tag	„19860302“

Normalisierung für das Attribut Geschlecht

Beispiel: „Weiblich“

Normalisierung	Beispiel
m: male, männlich f: female, weiblich x: divers o: other, sonstige u: unknown, unbekannt	„f“

Bloomfilter Testdaten

Für die Erzeugung der Testdaten wurde das Random-Hashing verwendet. Dabei wird ein Cryptographic Longterm Key erzeugt. Dies bedeutet, dass mehrere Attribute im selben Bloomfilter codiert werden. Dies stellt einen weiteren Schutzmechanismus dar. Der Bloomfilter besitzt eine Länge von 1000 Bits. Diese verdoppelt sich auf 2000 Bits, durch die Erzeugung von Balanced Bloomfiltern, was ein Verfahren zur Härtung von Bloomfilter darstellt. Pro Attribut (und zusätzlich für die Erzeugung von Balanced Bloomfilter) wird je ein Startwert (Seed) benötigt, der später bei allen Standorten (je Attribut) identisch sein muss. Dies stellt ein Geheimnis dar, welches nur den Standorten bekannt sein sollte. Für die Testdaten wurden folgende Startwerte/Seeds verwendet:

Vorname: 923764623

Nachname: 237845895

Geschlecht: 324566934

Geburtsdatum: 937457589

Postleitzahl: 476294534

Balanced Bloomfilter: 457943583

Die Bloomfilter wurden mit dem E-PIX erzeugt. Wenn Sie selbst Bloomfilter erzeugen möchten, können Sie die bereits vorbereitete Konfiguration nutzen, die bereitgestellt wird.

Es werden mehrere Testdatensätze bereitgestellt, um verschiedene Softwarelösungen zu testen. Die Datei mit den Testdaten, welche artifizielle Daten und den daraus resultierenden Bloomfilter beinhaltet, ist als CSV-Datei beigelegt.

Die Testdatensätze bestehen insgesamt aus 80 Datensätzen. Dabei gehören je zwei Datensätze zur selben (fiktiven) Person. Die erste Spalte der Testdaten gibt dabei an, welche Datensätze zusammengehören. Im späteren Verlauf des Projekts, sollen mehrere Konfigurationen mittels Echtdaten werden, die Testdaten dienen nur dazu, die eigene Software einzurichten und Ergebnisse miteinander abzugleichen.

Im Folgenden ist ein Testdatensatz dargestellt.

Attribut	Wert
Vorname	Lukas
Nachname	Müller
Geschlecht	Männlich
Geburtsdatum	14.03.1992
Resultierender Bloomfilter	Sn6HefKrSNtOLbsIfpSLwpi4OLxKCDGYI6o1yf6whjh62XsxWpe53ntCWawOe4ZOt UuMKwrLqzQPZtmn5Oo1wgz5+7jtqDIWpp1v+73BPMCg6qdMAmKQLF330I7hP D4GntSJllbLL6s/VJL0hVVhRc1EQWYHDAR2Useaqsw2duSneBdU00703n4rHs2f 3wCc4xOqeGuPixnF7kgMrxdBbj08/HDbi5tuif1g+4RCwULqfeZTz4yFvAsgj570nE MAAGwg6tn/OGOKGulokdRX6ZEGd1u4ZP0ybKttALEFWxyA7iVJwcPg1AYEQ/zAf vn2zSVoH/C2gbQjQ==