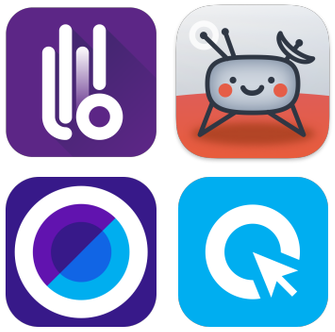
Analytics & Privacy

Gute Analytics und gute Privatsphäre – Geht das?

Daniel Jilg, Web&Wine, Februar 2021



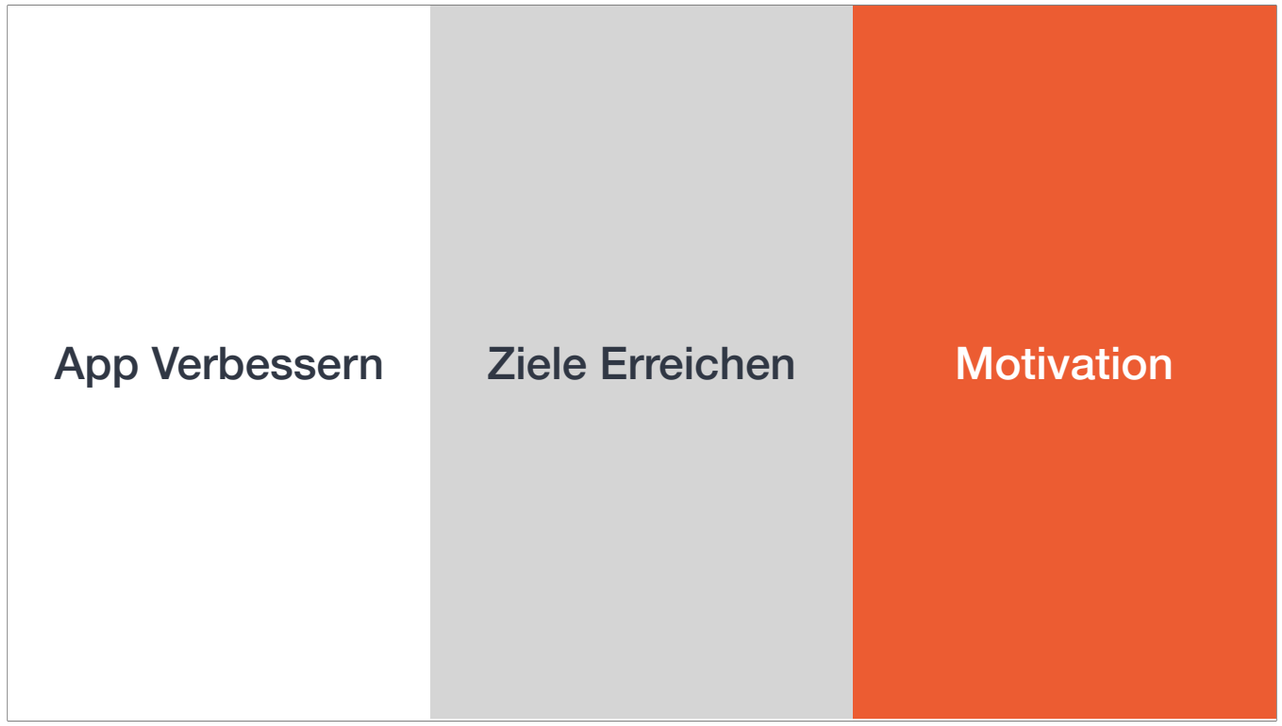
Daniel Jilg

Twitter: breakthesystem
Github: winsmith

- Findet Statistiken und quantification cool
- Mag Graphen
- Mag Privacy
- Hat Privacy-Aware Firmen wie Keepsafe und Cliqz gearbeitet

- Für beide: Browser gebaut die die Privatsphäre der Benutzer schützen

Warum Analytics?



- Nur wenn ich weiß wie meine App benutzt wird, kann ich sie effektiv verbessern
- Wenn ich bestimmte Ziele mit meiner App erreichen will (Verkäufe, glückliche User*innen, Sign-Ups, Userzahlen) erreiche ich sie eher wenn ich den Weg dorthin analysiere und optimiere
- Es motiviert mich tatsächlich einfach SEHR, zu sehen wie Leute meine App benutzen

GDPR & CCPA

Gesetzestexte yay! 😍🍪

Diese Präsentation benutzt Cookies bitte Cookies erlauben weil Cookies sind wichtig und so weiter mit dem Betrachten dieser Präsentation erlauben Sie dem Autor mindestens einen Cookie zu essen möglicherweise auch mehr das kann man nicht so genau

- Okay, wir haben uns also entschieden, Daten zu sammeln.
- Müssen also die DSGVO beachten
- DSGVP, auch GDPR genannt, und das Kalifornische Pendant dazu, die CCPA (California Consumer Privacy Act)
- das sind die Gesetze die Benutzer*innen vor dem Missbrauch ihrer Daten schützen sollen
- Benutzer*innen sollen Kontrolle über ihre Daten haben
- Führt unter anderem zu diesen schrecklichen Cookie consent Bannern

„personenbezogene Daten“ [bezeichnet] alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person […] beziehen;

Art. 4 DSGVO

Wichtigstes Kriterium:

- Personenbezogene Daten
- Englisch: Personally Identifiable Information (PII)
- Sobald Personenbezogene Daten gespeichert werden, hat die entsprechende Person wichtige Rechte:
    - Recht auf Auskunft
    - Recht auf Berichtigung
    - Recht auf Löschung
- Das sind alles wichtige Rechte, aber wir wollen ja gar nicht über einzelne Personen Bescheid wissen! Wir wollen wissen, wie viel Prozent unserer User den In App Purchase screen angesehen haben!

als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen identifiziert werden kann, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind;

Art. 4 DSGVO

Als identifizierbar gilt, wer identifiziert werden kann, also wer erkannt werden kann…

als identifizierbar wird eine natürliche Person angesehen, die direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie einem Namen, zu einer Kennnummer, zu Standortdaten, zu einer Online-Kennung oder zu einem oder mehreren besonderen Merkmalen identifiziert werden kann, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind;

Art. 4 DSGVO

… und zwar vor Allem anhand von IDs, Namen, Standortdaten oder besonderen Merkmalen.

Diese Sachen wollen wir also eigentlich NICHT speichern

„Pseudonymisierung“ [bezeichnet] **die Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können,** sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden;

Art. 4 DSGVO

Was für Optionen haben wir? Eine ist die Pseudonymisierung…

„Pseudonymisierung“ [bezeichnet] die Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden;

Art. 4 DSGVO

Wenn wir unsere Daten Pseudonymisieren können sie nicht mehr direkt spezifischen personen zugeordnet werden

„Pseudonymisierung“ [bezeichnet] die Verarbeitung personenbezogener Daten in einer Weise, dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden;

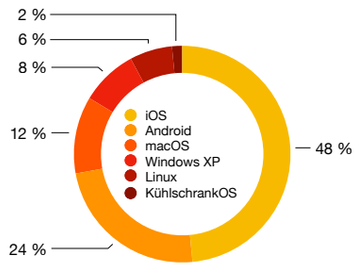
Art. 4 DSGVO

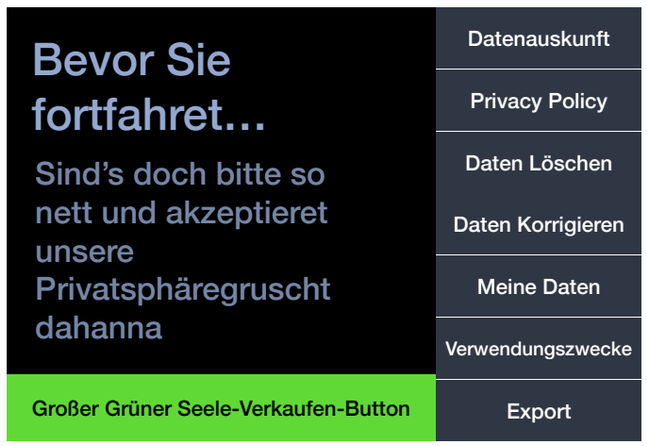
- Pseudonymisieren bedeutet aber auch, dass Daten nicht völlig anonym sind
- wenn jemand ZUSÄTZLICHE Informationen zur Verfügung hat
- kann durch Zusammenführen der Daten wieder auf spezifische Personen kommen
- Fällt also immer noch unter die DSGVO
- Also immer noch Consent Dialoge, und Infrastruktur zur Verwaltung und Löschung von Daten

Die Grundsätze des Datenschutzes sollten daher nicht für anonyme Informationen gelten, d.h. für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder personenbezogene Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann.

Erwägungsgrund 26, Satz 5, DSGVO

Die einzige Option, nicht unter die DSGVO zu fallen ist
- daten so anonymisierten
- Dass sie nicht mehr auf personen zurück zu führen sind





- Was heißt das für Analytics?
    - Wir wollen statistische Daten über die Benutzung unserer Apps und Webseiten
    - Wir wollen KEINE User Consent Dialoge, weil die einfach super nervig sind
    - Und wir wollen erst Recht nicht in die Situation kommen, Personenbezogene Daten speichern zu müssen und dann die gesamte Infrastruktur aufbauen müssen, um damit verantwortungsbewusst umgehen
    - Personen haben Rechte, über ihre personenbezogenen Daten zu bestimmen
        - User Choice vor dem Speichern
        - Datenauskunft
        - Detaillierte Privacy Policy
        - Recht auf Löschung
        - Recht auf Berichtigung
        - Recht auf Auskunft
        - Recht auf Verarbeitungseinschränkung der Daten
        - und Recht auf Datenübertragbarkeit
- Wir wollen diese Infrastruktur nicht aufbauen!

- Wir wollen anonymisierte statistische Daten!
- Die können uns die wichtigen Fragen beantworten
- Personenbezogene Daten sind wie radioaktiver Müll
- Den wollen wir gar nicht erst sammeln

Das waren die Grundüberlegungen

Google Analytics & Firebase

Data Processing Agreement ✅

Data Deletion Requests ✅

Data Retention Settings 🆗

Unique ID across all Google Services 😱

Google uses visitor data for Ads, Youtube, AdSense, and maybe more 🤖

Needs Consent Dialog 👻

App Store Analytics Warning 🙄

Not Stored in Europe ❌

Hard to write Privacy Policy 😭

- Die größten Anbieter im Bereich analytics und App analytics:
- Google Analytics, Google Analytics 360 und Firebase
- gehören alle Google
- Möchten wir nicht benutzen aus diesen Gründen

Strategien zur Lösung

Wir wissen welche Balance wir brauchen, was sind Strategien und Aspekte einer idealen lösung?

Datensparsamkeit

https://www.datenschutz.org/datensparsamkeit/
https://martinfowler.com/bliki/Datensparsamkeit.html

So wenige Daten wie möglich speichern

Anonyme (hashed) user IDs

Pro App eigene User IDs

https://gdpr-info.eu/recitals/no-26/

Keine User Identifier direkt Speichern, wenn dann nur hashed, und für jede App andere

Session Based IDs

https://piwik.pro/blog/how-to-do-useful-analytics-without-personal-data/

- Noch besser als User Identifier:
- Session Identifier.
- Jede Session (also jeder app Launch, jeder website Aufruf) bekommt eine neue User ID
- Erlaubt kein direktes Zählen von Usern, aber Flow durch die App analysieren
- Retention, und analyse von Dimensionen durch metadaten

differential privacy

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/dwork.pdf
https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf
https://www.cerias.purdue.edu/news_and_events/events/security_seminar/details/index/j9cvs3as2h1qds1jrdqfdc3hu8

- kalkuliertes einsprühen von zufälligen Falschdaten
- Diese sind so verteilt, dass sie sich statistisch gesehen bei grossen Datensätzen wieder rausrechnen
- Statistische Analyse Möglich, aber einzelne Datensätze sind nicht personalisiert

unsafe data removal

https://cliqz.com/en/magazine/how-we-at-cliqz-protect-users-from-web-tracking
https://www.researchgate.net/publication/312638031_Tracking_the_Trackers
https://www.researchgate.net/publication/334316330_Privacy-Preserving_Classification_with_Secret_Vector_Machines
https://cliqz.com/en/magazine/how-we-at-cliqz-protect-users-from-web-tracking

- Bei Cliqz gemacht
- Alle key-value-pairs in einen cache
- Neue values werden erst gespeichert wenn sie bei mindestens *n* verschiedenen Leuten aufgetaucht sind
- So kann verhindert werden dass personen durch die Values identifiziert werden können

Analyse the URL, headers and postdata of the request.
Tokenise this data into key-value pairs.
Evaluate the safeness of each key-value pair.
If there are unsafe values, remove the data from the request.

Die Daniel-Methode: Telemetry

- hab mir das alles zu Herzen genommen
- Mein System heisst Telemetry — Telemetrie sind die Betriebsdaten die zb Raumschiffe nach Hause schicken
- Überblick wie ich Analytics mache
- Sowohl für apps als auch für Websites

User based IDs

Install based IDs

Session IDs

Keine IDs

Was für IDs? Das system nimmt diese hier

Events heißen bei mir Signals

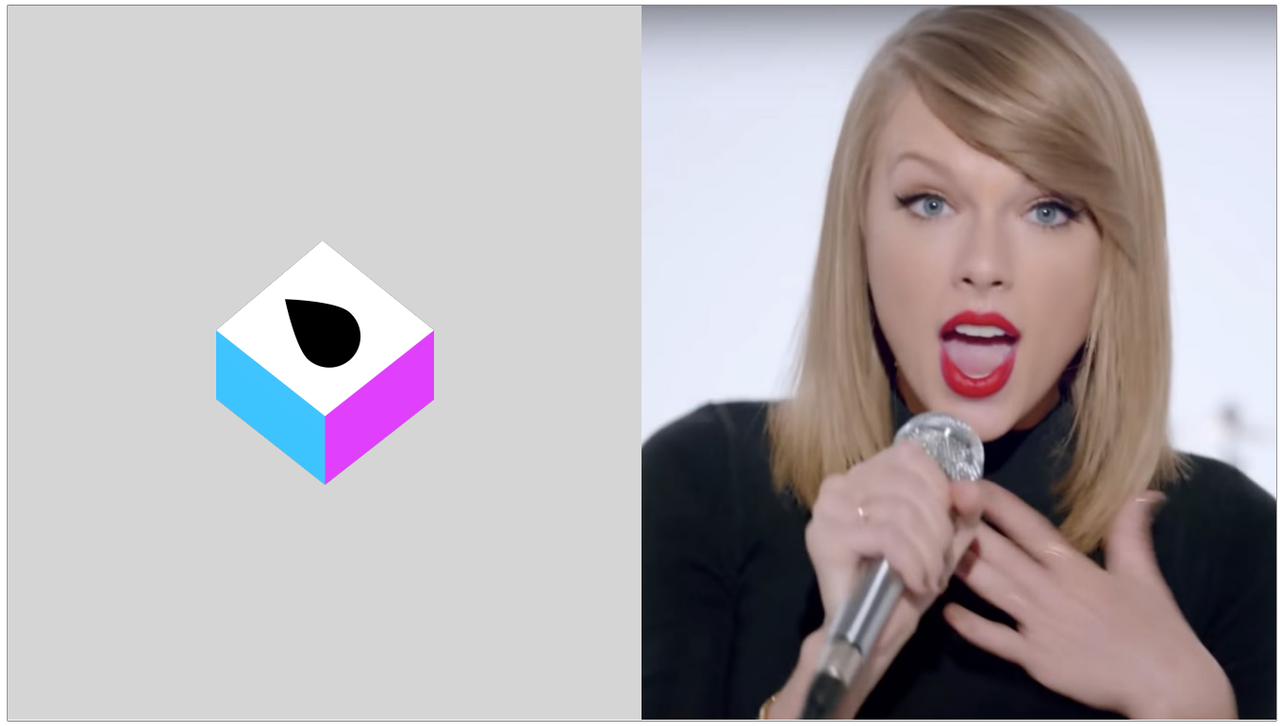
Signals mit type, metadata payload und hashed id

Meine Apps und Webseiten schicken Telemetrie-Signale an meinen Server, die bestehen aus

- Type (was ist passiert?)
- App ID (um welche App geht es)
- Metadata Payload (beliebige Randdaten als Dictionary)
- Hashed ID (so dass ich nie mehr die ursprüngliche ID errechnen kann — Damit anonym

```
{
   "clientUser": "3ef57dcb4728bdc726b93b383a704712ae2981196cb47794a19f2b787e9c1616",
   "receivedAt": "2021-02-12T13:25:48+0000",
   "appID": "BD342A23-826F-4390-BC0F-7CD34A5CE7F8",
   "type": "PauseAction",
   "payload": {
      "isTestFlight": "false",
      "platform": "iOS",
      "isSimulator": "false",
      "isAppStore": "true",
      "buildNumber": "2",
      "modelName": "iPhone11,2",
      "appVersion": "3.11.0",
      "systemVersion": "iOS  14.4”
   }
}
```

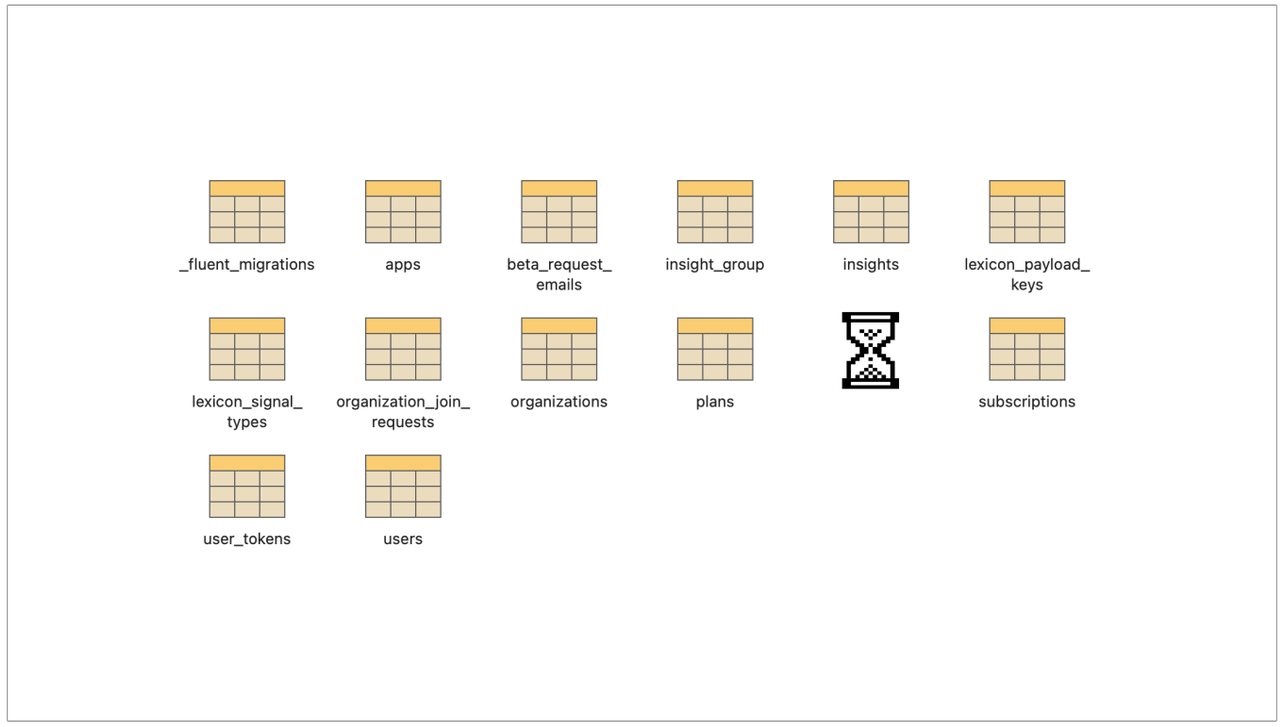

Das ist ein echtes Signal aus meiner Datenbank, das hab ich aus dem Strom gefischt als ich diese Folie gemacht hab



Der server ist in Swift geschrieben und benutzt das Vapor-Framework

Daten speichere ich in PostgreSQL dazu kommen seit kurzem auch Kafka und Druid, dazu gleich mehr

- Die postgres Datenbank sieht so aus wie man sich das vorstellt, keine großen Überraschungen
- Aber es stellt sich raus, ab so ca 9 Millionen Signalen



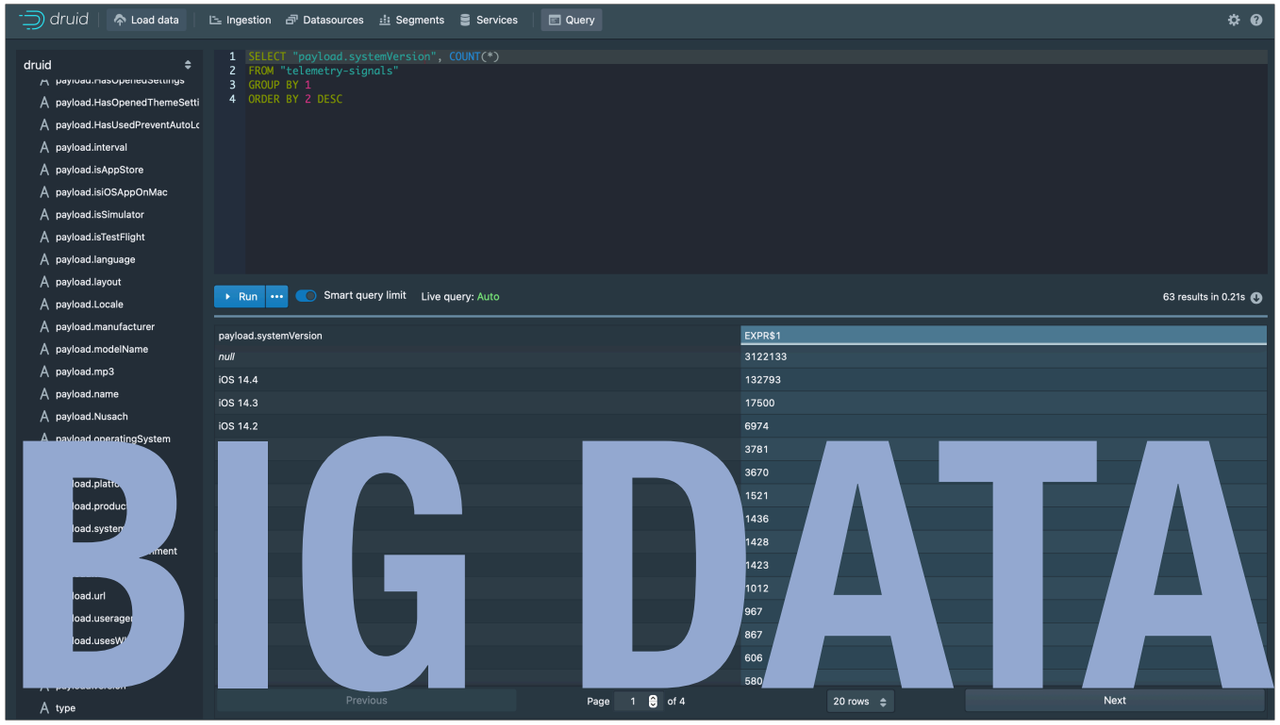
- Wird es sehr sehr langsam die statistisch zu analysieren

HyperLogLog

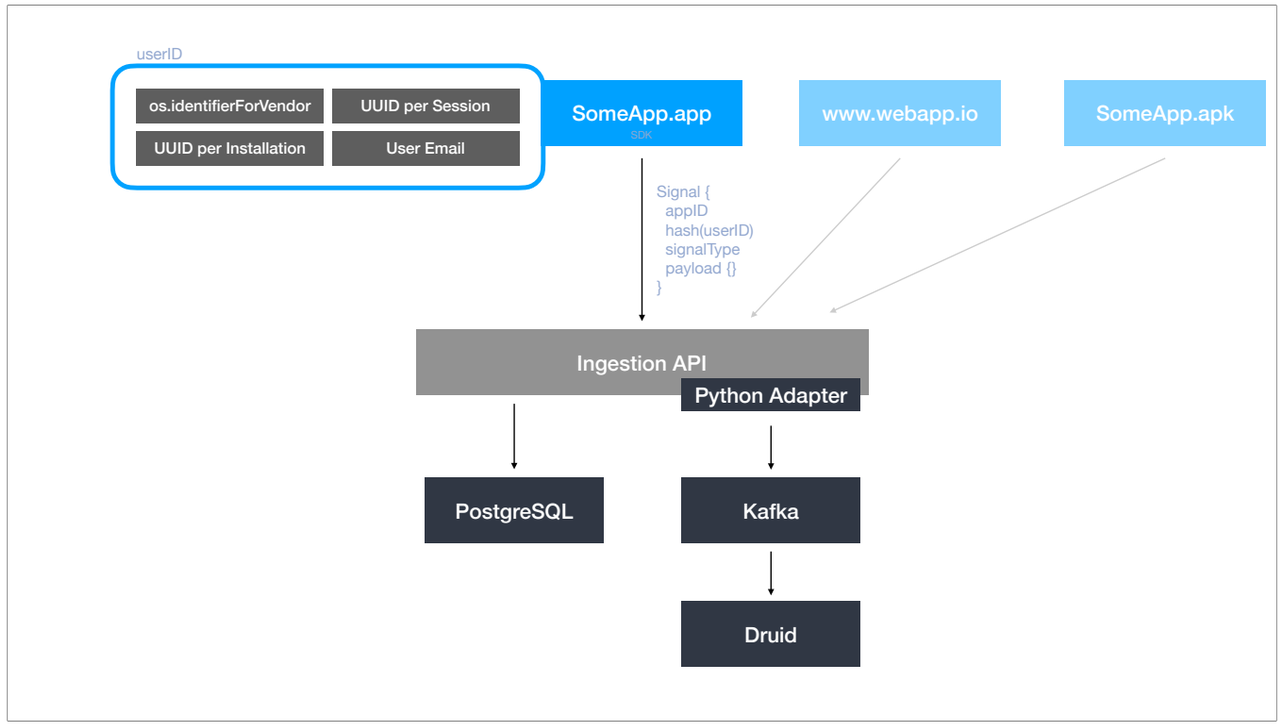
http://algo.inria.fr/flajolet/Publications/FlFuGaMe07.pdf
http://algo.inria.fr/flajolet/Publications/DuFl03-LNCS.pdf
http://algo.inria.fr/flajolet/Publications/FlMa85.pdf

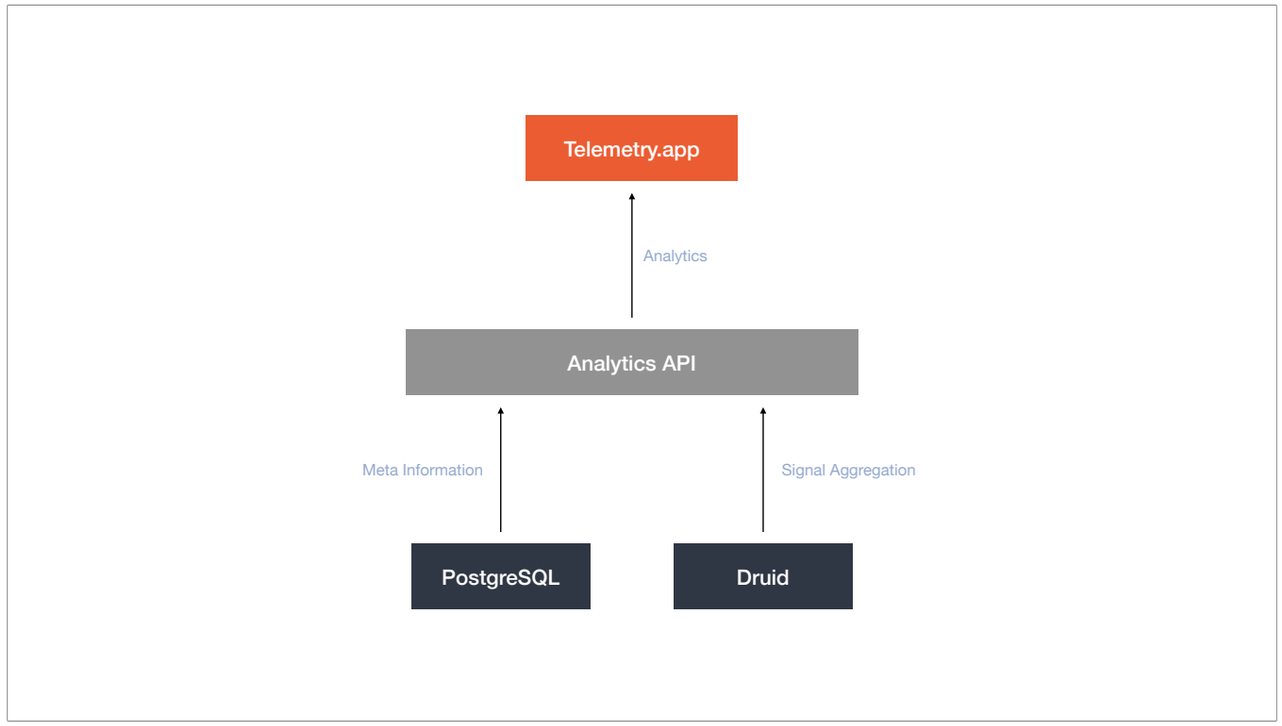
Die Lösung dafür sind entweder

- Nur einmal am Tag die Berechnung laufen lassen (laangweilig)
- Approximationsalgorithmen wie HyperLogLog mit denen sich statistische Zählungen super schnell durchführen lassen
- Die sind halt dann ungenau, wenn ich 1000 User*innen habe, sehe ich manchmal 1001 oder 999
- Rechnet sich ähnlich wie bei differential privacy in der masse wieder raus

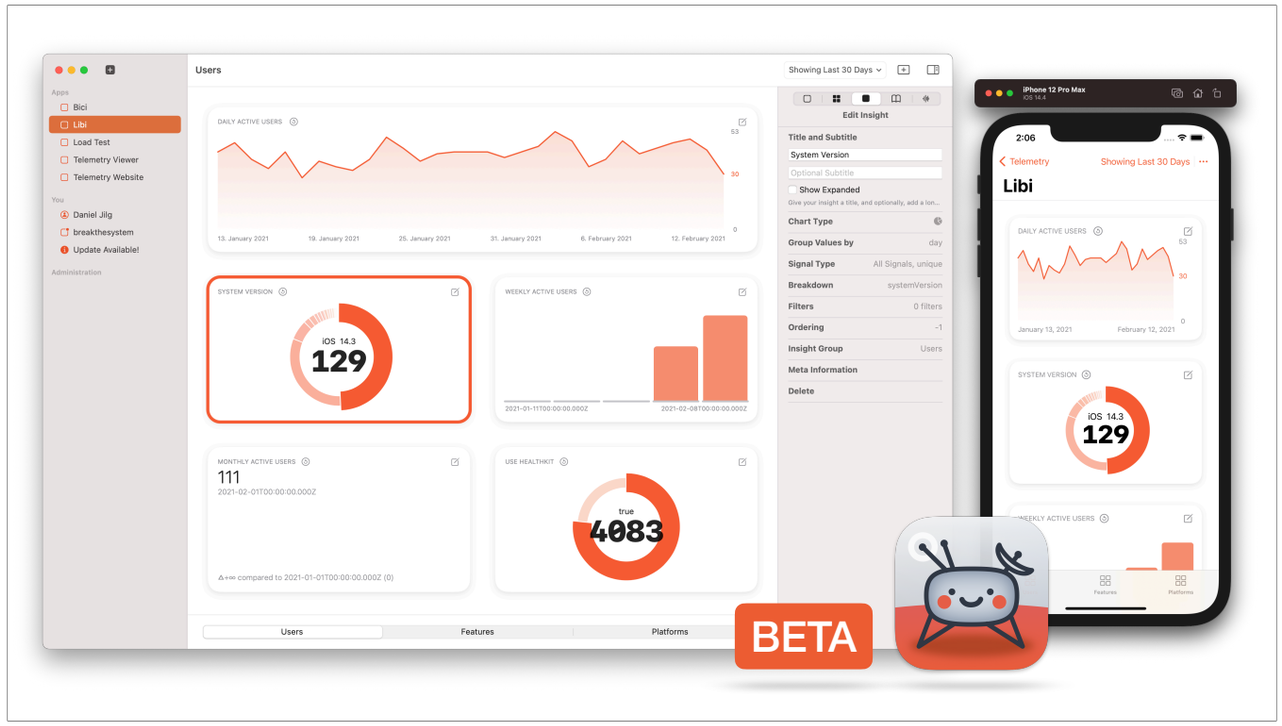


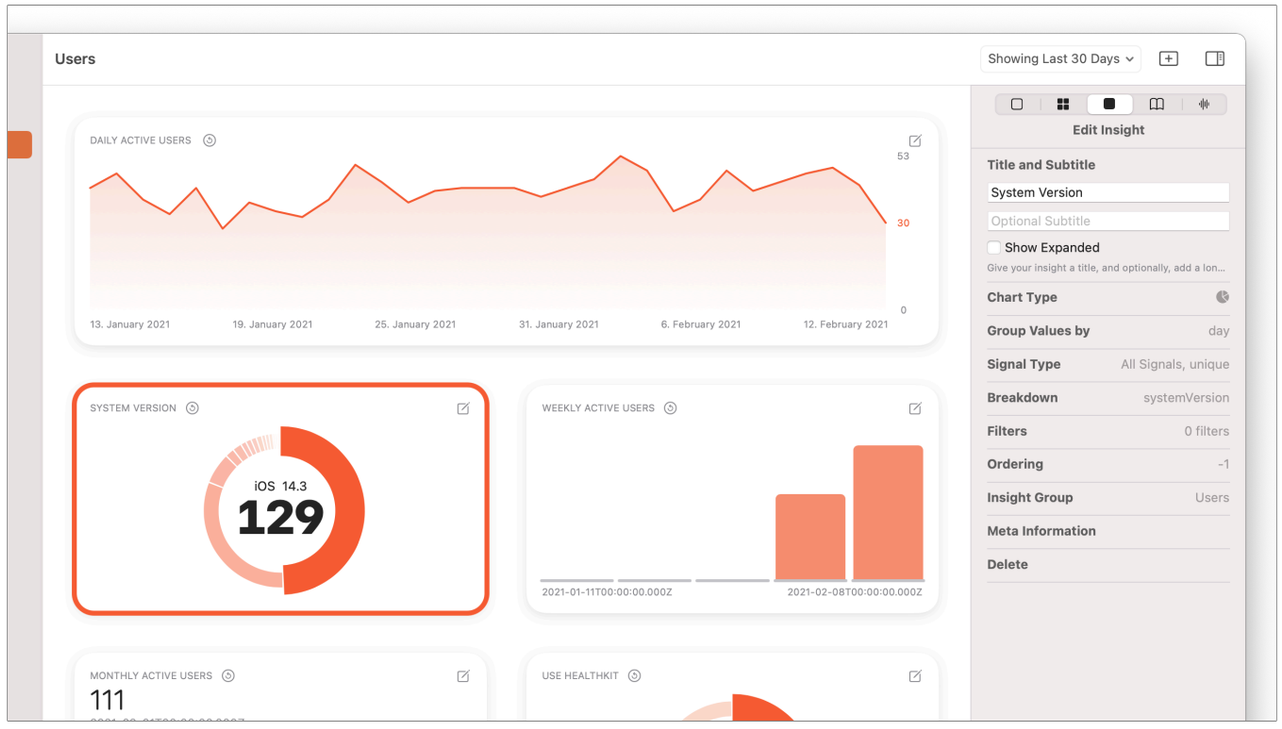
- Die Algorithmen implementier ich nicht selber
- sondern benutze Kafka, einen Distributed Event Streaming Server,
- und Druid, einen Data Lake und Real Time Analytics Database Server.
- Ich äh betreibe also BIG DATA
- Damit kann ich hier zb ca 9 Millionen Signale in 210 Millisekunden zu Betriebssystemen gruppieren

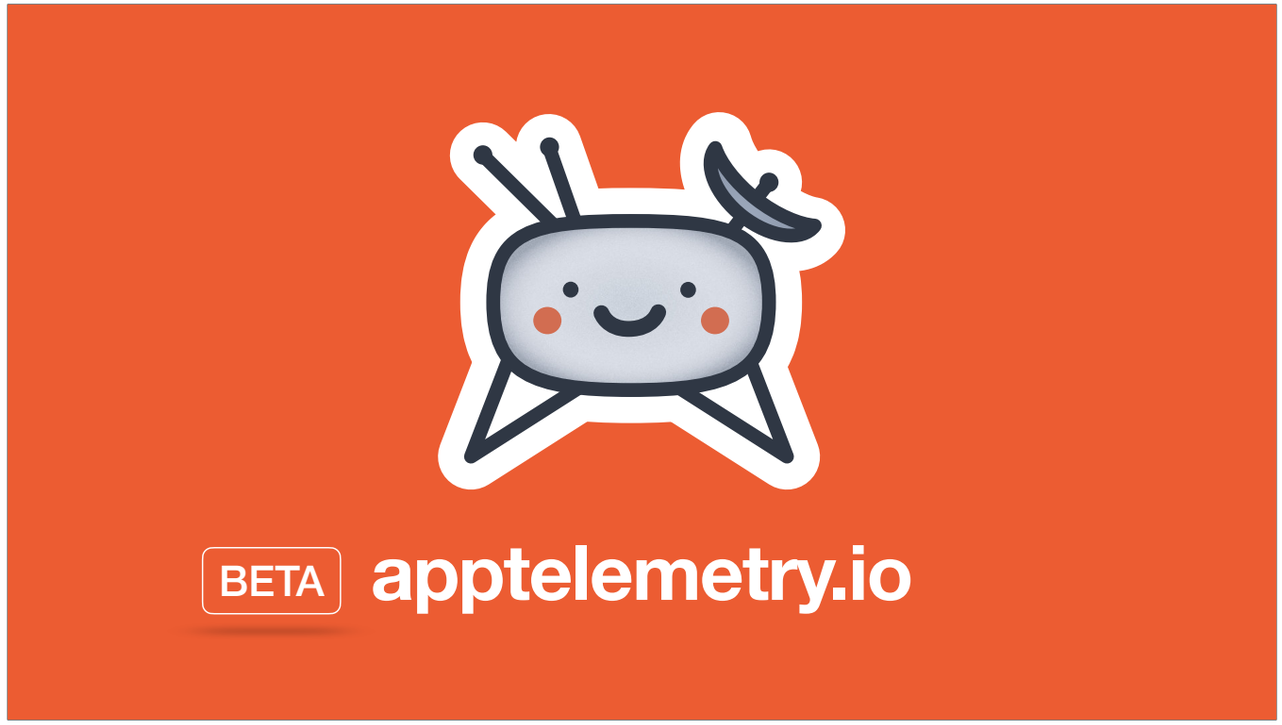

userID
os.identifierForVendor
UUID per Session
UUID per Installation
User Email
SomeApp.app
SDK
www.webapp.io
SomeApp.apk
Signal {
appID
hash(userID)
signalType
payload {}
}
Ingestion API
Python Adapter
PostgreSQL
Kafka
Druid


Telemetry.app
Analytics
Analytics API
Meta Information
Signal Aggregation
PostgreSQL
Druid

Telemetry


Users
Showing Last 30 Days
Apps
Bici
Libi
Load Test
Telemetry Viewer
Telemetry Website
You
Daniel Jilg
breakthesystem
Update Available!
Administration
DAILY ACTIVE USERS
53
30
0
13. January 2021
19. January 2021
25. January 2021
31. January 2021
6. February 2021
12. February 2021
SYSTEM VERSION
iOS 14.3
129
WEEKLY ACTIVE USERS
2021-01-11T00:00:00.000Z
2021-02-08T00:00:00.000Z
MONTHLY ACTIVE USERS
111
2021-02-01T00:00:00.000Z
compared to 2021-01-01T00:00:00.000Z (0)
USE HEALTHKIT
true
4083
Users
Features
Platforms
Edit Insight
Title and Subtitle
System Version
Optional Subtitle
Show Expanded
Give your insight a title, and optionally, add a lon...
Chart Type
Group Values by
day
Signal Type
All Signals, unique
Breakdown
systemVersion
Filters
0 filters
Ordering
-1
Insight Group
Users
Meta Information
Delete
iPhone 12 Pro Max
iOS 14.4
2:06
Telemetry
Showing Last 30 Days
Libi
DAILY ACTIVE USERS
53
30
0
January 13, 2021
February 12, 2021
SYSTEM VERSION
iOS 14.3
129
WEEKLY ACTIVE USERS
Features
Platforms
BETA


Users
Showing Last 30 Days
DAILY ACTIVE USERS
53
30
0
13. January 2021
19. January 2021
25. January 2021
31. January 2021
6. February 2021
12. February 2021
SYSTEM VERSION
iOS 14.3
129
WEEKLY ACTIVE USERS
2021-01-11T00:00:00.000Z
2021-02-08T00:00:00.000Z
MONTHLY ACTIVE USERS
111
USE HEALTHKIT
Edit Insight
Title and Subtitle
System Version
Optional Subtitle
Show Expanded
Give your insight a title, and optionally, add a lon...
Chart Type
Group Values by
day
Signal Type
All Signals, unique
Breakdown
systemVersion
Filters
0 filters
Ordering
-1
Insight Group
Users
Meta Information
Delete


BETA
apptelemetry.io

Daniel Jilg

Twitter: breakthesystem
Github: winsmith
apptelemetry.io