Analytics Pioneers

COMMUNITY TRAININGS

Daten-Lifecycles in BigQuery meistern
- PII löschen, Insights bewahren -

Von der Community für die Community

Analytics Pioneers Community wurde gegründet, um den regelmäßigen Austausch von Ideen, Herausforderungen und Lösungen in der Digital Analytics Community zu fördern.

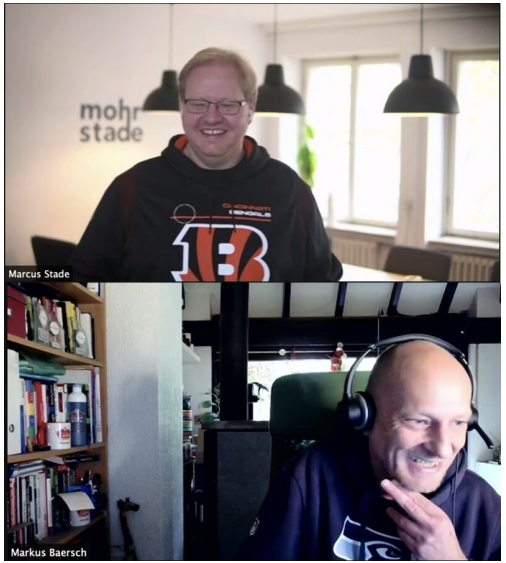
Wir wollen die Analytics Community in Europa näher zusammenbringen, vernetzen und eine Plattform bieten, auf der Mitglieder voneinander lernen und Spaß haben.

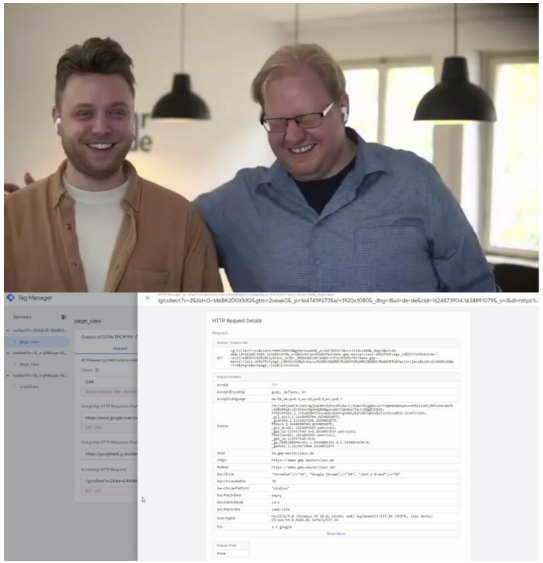
Jeder darf und soll sich einbringen!

Analytics Pioneers Community

Meetups

Community Trainings





www.analytics-pioneers.com

Analytics Pioneers

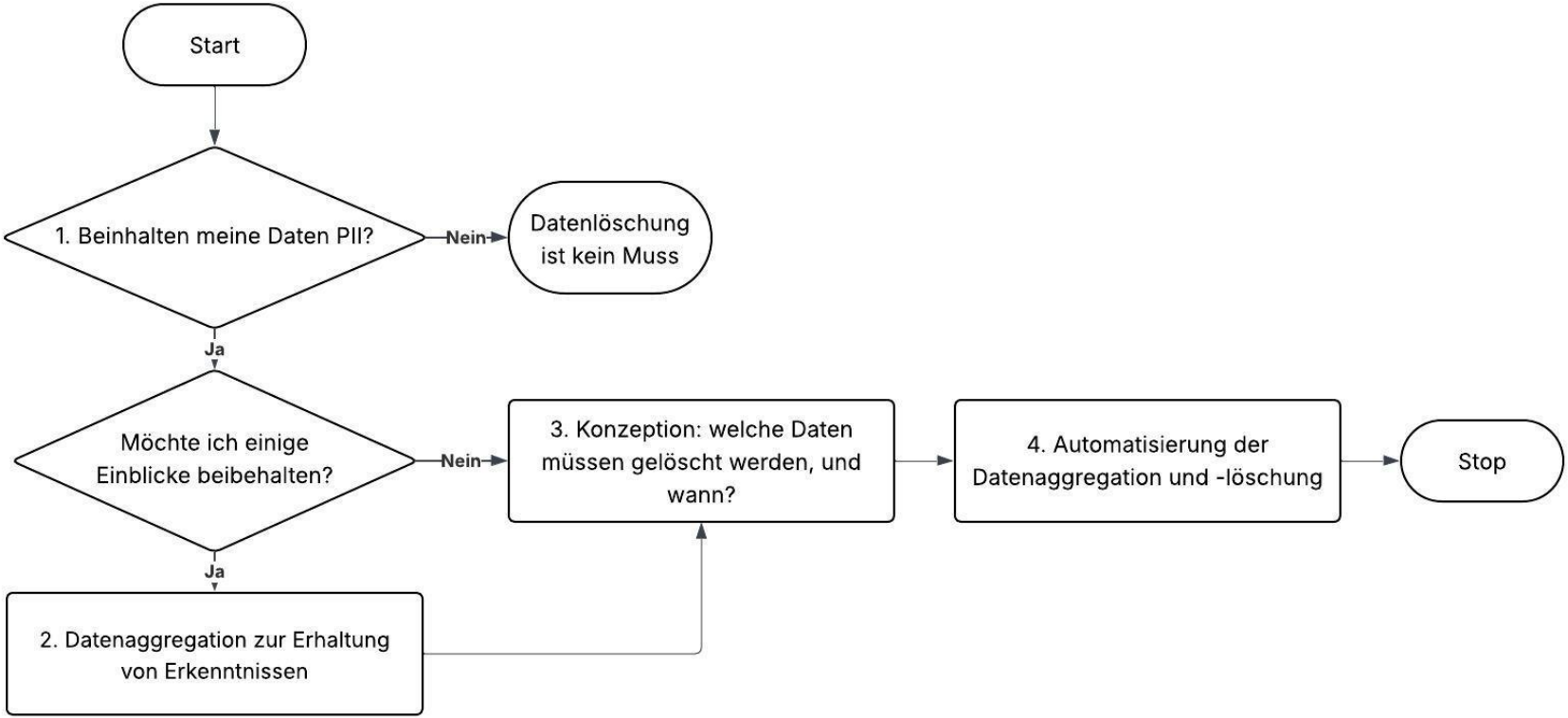
Die Analytics Pioneers finden bereits in 37 Städten statt

Ihr findet die Slides & Code Beispiele zu unseren Trainings immer hier:

https://gitlab.com/mstade81/analytics-pioneers-community-trainings

Analytics Pioneers

Agenda



Analytics Pioneers

Agenda

1. Enthalten meine Daten PII?
2. Daten Aggregieren zur Erhaltung von Erkenntnissen
3. Konzeption: welche Daten müssen gelöscht werden, und wann?
4. Automatisierung der Datenaggregation und -löschung

Analytics Pioneers

Agenda

1. **Enthalten meine Daten PII?**
2. Daten Aggregieren zur Erhaltung von Erkenntnissen
3. Konzeption: welche Daten müssen gelöscht werden, und wann?
4. Automatisierung der Datenaggregation und -löschung

Disclaimer

Diese Präsentation dient ausschließlich zu Informationszwecken und stellt **keine rechtliche Beratung** dar. Wir teilen hier technische Best Practices, jedoch keine rechtlichen Empfehlungen. Für rechtliche Fragen oder Unsicherheiten empfehlen wir, einen **qualifizierten Rechtsberater** zu konsultieren.

Definition von personenbezogenen Daten

Personenbezogene Daten (PII) umfassen alle Informationen, die sich auf eine identifizierte oder identifizierbare **natürliche Person** beziehen. Dazu gehören:

1. **Direkte Identifikatoren** wie Name, Adresse, E-Mail oder Telefonnummer.
2. **Indirekte Identifikatoren** wie IP-Adresse, Cookies, Standortdaten oder Geräte-IDs.
3. **Besonders sensible Daten**, wie Gesundheitsinformationen, biometrische oder genetische Daten, die zusätzlichen Datenschutzanforderungen unterliegen.

Analytics Pioneers

Entdecken von PII in gängigen Datenquellen

Gängige Datenquellen, die PII enthalten, umfassen:

- **CRM-Daten** (user_id, E-Mail, Name, Telefonnummer, Adresse usw.) von Hubspot, Salesforce usw.
- **Google Analytics-Rohdaten** (user_id, geografische Standortdaten, Geräteinfos usw.): analytics_********
- **Google Ads-gclid-Infos**: ads_ClickStats
- **Backend-Daten**

Analytics Pioneers

Agenda

1. Enthalten meine Daten PII?
2. **Daten Aggregieren zur Erhaltung von Erkenntnissen**
3. Konzeption: welche Daten müssen gelöscht werden, und wann?
4. Automatisierung der Datenaggregation und -löschung

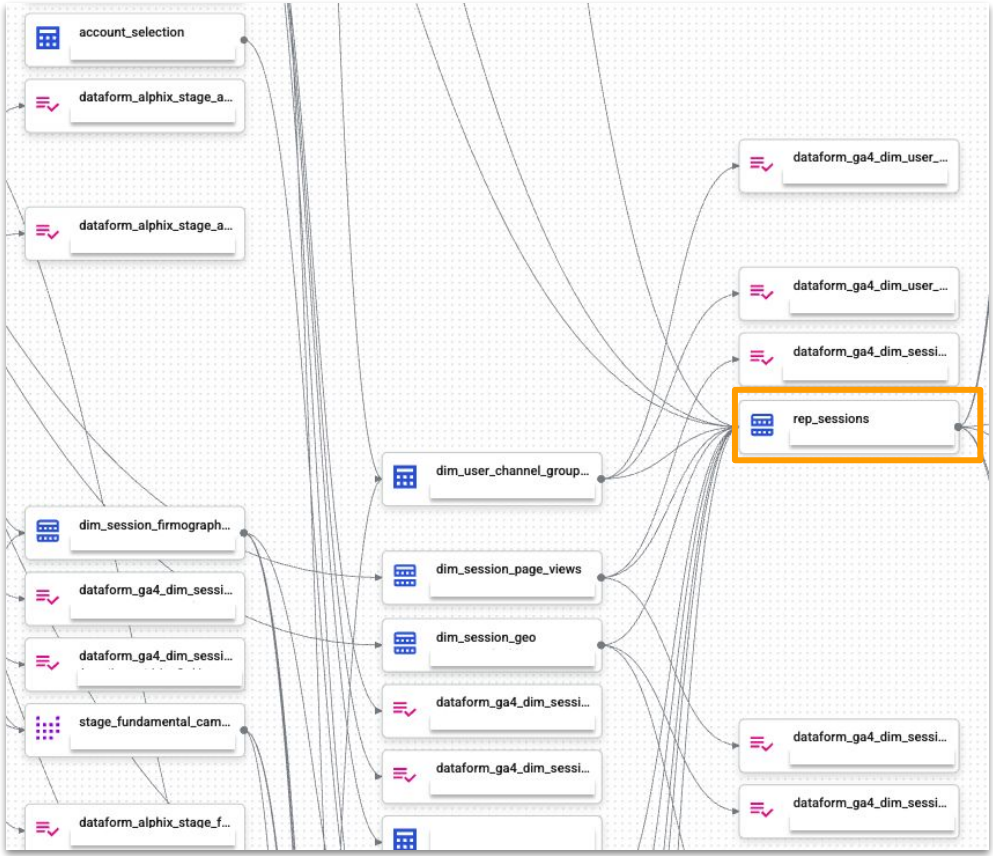
Welche Erkenntnisse müssen aufbewahrt werden?

Nach zwei Jahren sind keine Sitzungs-, Kunden- oder sogar Klick-ID-Daten mehr vorhanden – kein Problem! Daten von solcher Granularität sind nach dieser Zeit wahrscheinlich ohnehin **nicht mehr geschäftsrelevant**.

Es könnte jedoch relevant sein, die Trends beizubehalten:

- **Geografischer Standort**, z.B. "Wie hat sich der Traffic in Osteuropa im Laufe der Zeit entwickelt?"
- **Geräteinfos**, z.B. "Wie hat sich der Anteil des iOS-Traffics im Vergleich zu vor zwei Jahren verändert?"
- **Attribution**, z.B. "Welche Kanäle, die zuvor nicht relevant waren, sind es jetzt?"

Wichtige Modelle aggregieren vor Löschung der PII-Daten



"Finale" Tabellen in der Pipeline identifizieren. **Enthalten sie PII-Daten?** Wenn ja -> aggregieren!

-> für Details siehe unsere Github-Repo

Tricks für die Aggregationstabellen

```
rowConditions: [
    "number_of_sessions > 4",
    "number_of_users > 4"
]
```

```
HAVING
    number_of_sessions > 4
AND number_of_users > 4
```

- **K-Anonymität!** Wir berichten nicht über Kategorien, die

weniger als eine bestimmte Anzahl von Sitzungen/Nutzern haben (z.B. nur ein Nutzer auf unserer Website am 1. Januar, der in Magdeburg wohnt).

- **Wähle deine GROUP BY-Klausel mit Bedacht.**

Wenn deine Kategorien zu spezifisch sind, führt die K-Anonymität zu einem großen Datenverlust.

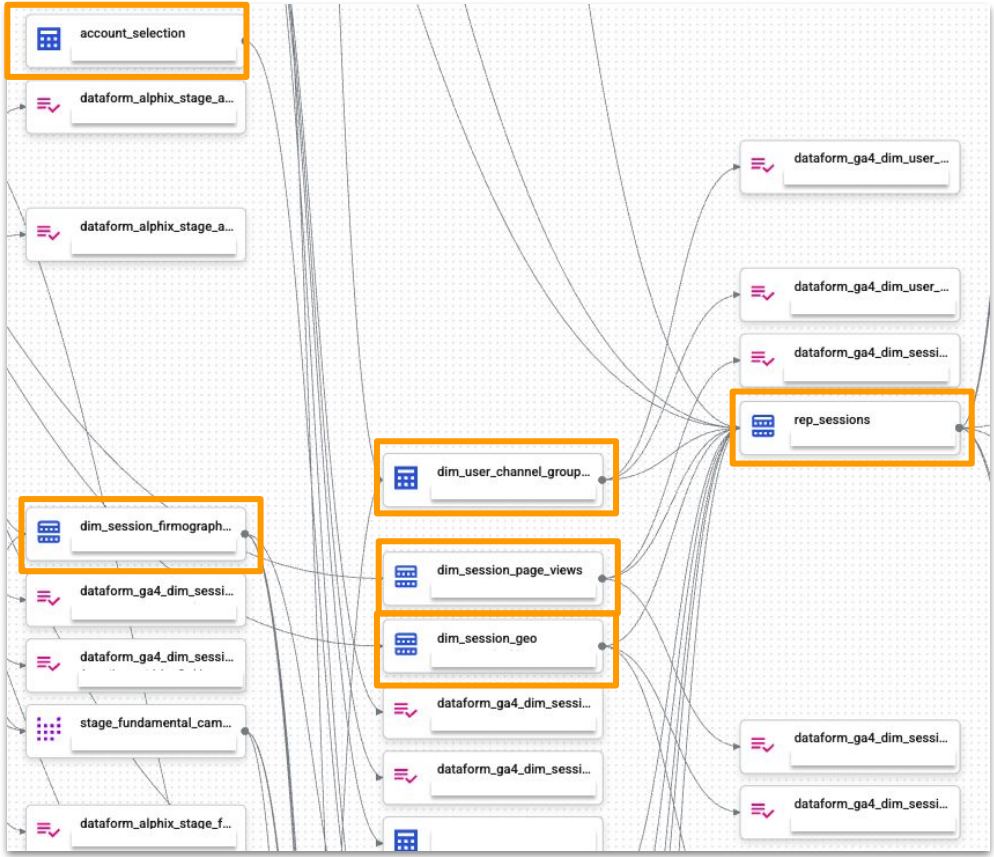
- **Wenn die tägliche Aggregation zu Datenverlusten führt** (wieder aufgrund

von K-Anonymität), verwende stattdessen eine wöchentliche oder monatliche Aggregation (z.B. session_week, session_month).

Agenda

1. Enthalten meine Daten PII?
2. Daten Aggregieren zur Erhaltung von Erkenntnissen
3. **Konzeption: welche Daten müssen gelöscht werden, und wann?**
4. Automatisierung der Datenaggregation und -löschung

Welche Daten werden gelöscht und wann?



- Alle Tabellen, die **PII** enthalten
- Maximale Lebensdauer-Richtlinie: **2 Jahre**
- Abhängig von der Datenschutzrichtlinie: 2 Jahre seit Speicherung des Eintrags VS 2 Jahre **seit letzter Interaktion mit dem Nutzer**

Agenda

1. Enthalten meine Daten PII?
2. Daten Aggregieren zur Erhaltung von Erkenntnissen
3. Konzeption: welche Daten müssen gelöscht werden, und wann?
4. **Automatisierung der Datenaggregation und -löschung**

Datenaggregation automatisieren

- Einen separaten Workflow erstellen oder die aggregierten Tabellen in einen bestehenden Workflow integrieren, der **regelmäßig ausgeführt** wird.
- **Tabellenpartitionierung und Inkrementalität** nutzen, um Rechenkosten zu sparen.
- Eine **versehentliche Löschung** der aggregierten Tabellen durch die Verwendung des Keywords `protected:true` verhindern.

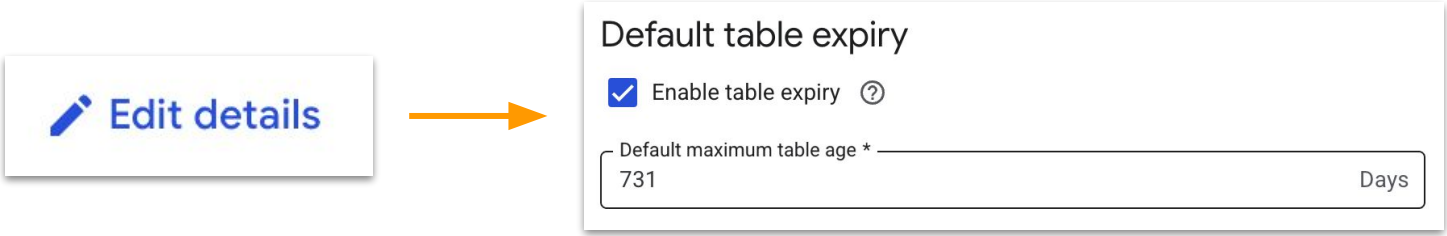
Datenlöschung automatisieren

- **Tabellenpartitionierung** in allen PII-haltigen Tabellen verwenden
- Nutze die **Partitionsablaufzeit-Funktion**. Du kannst sie direkt in Dataform oder mit SQL-Abfragen festlegen.

```
ALTER TABLE projekt.dataset.table
SET OPTIONS ( partition_expiration_days = 731 );
```

Table type	Partitioned
Partitioned by	DAY
Partitioned on field	session_date
Partition expiry	731 days
Partition filter	Not required

- Für Tabellen, die nur Daten für einen Tag enthalten, verwende stattdessen die **Tabellenablaufzeit**. Zum Beispiel bei den Rohdaten von GA4:



Analytics Pioneers

Bonus: Einmalige Datenlöschung für GA4-Rohdaten

- Nachdem die Ablaufzeit für alle neuen GA4-Tabellen festgelegt wurde, erhalten **nur die neuen Tabellen** sie automatisch
- Alle bestehenden Tabellen im GA4-Datensatz **müssen geändert werden**, um ein Ablaufdatum hinzuzufügen.
- Tabellen, die älter als 2 Jahre sind, können direkt gelöscht werden (mit einem SQL-Statement). Für die anderen ~731 Tabellen muss ein **Tabellenablaufdatum** festgelegt werden.
- Es ist möglich, die Ablaufdaten einzeln in der **Benutzeroberfläche** festzulegen, aber dies ist ein langwieriger und fehleranfälliger Prozess.
- Um die Ablaufdaten einfach und fehlerfrei iterativ festzulegen, kann man ein **Python-Skript** in einem BQ-Notebook oder ein **SQL-Statement** verwenden.

Eure Speaker

Moritz Bauer
Director Marketing Technology
Follow me on LinkedIn

Marina Sukhanova
Consultant Marketing Technology
Follow me on LinkedIn

Follow us on LinkedIn