Michael Janssen

Ist BigQuery die Lösung aller GA4-Analyseprobleme?

Im Gegensatz zu seinem Vorgänger Universal Analytics bietet Google Analytics 4 einen kostenlosen Export der Daten für BigQuery an. Dabei wird BigQuery von vielen als die ultimative Lösung für alle Probleme bei der Auswertung von Daten mit GA4 angepriesen. Doch ist es wirklich die Allzweckwaffe, die es zu sein verspricht?

Während in der Vergangenheit der Datenexport bei Google Analytics bisher nur den zahlenden Enterprise-Kunden vorbehalten war, hat sich das mit der Einführung von Google Analytics 4 geändert. Denn jetzt kommen auch die Nutzenden der kostenfreien Version in den Genuss dieses Features. Mit nur wenigen Klicks kann BigQuery mit GA4-Property verknüpft werden und die Daten werden dauerhaft exportiert und in BigQuery bereitgestellt.

In diesem Artikel geht es aber nicht um eine Anleitung, wie die Daten aufbereitet und genutzt werden können, sondern um die strategischen und technischen Anforderungen, damit die Daten überhaupt genutzt werden können. Denn auch wenn die Integration in BigQuery scheinbar leicht von der Hand geht, gibt es vieles zu wissen, zu bedenken und zu entscheiden.

Die Liste an möglichen Vorteilen und Nutzungsmöglichkeiten der Daten in BigQuery ist groß. Aber besonders sind folgende Vorteile hervorzuheben:

Vorteil eins: längere Speicherdauer

In seiner kostenfreien Variante bewahrt GA4 Nutzer- und Ereignisdaten bis zu einem Zeitraum von maximal 14 Monaten auf. Nach Ablauf dieser Frist erfolgt eine automatische Löschung dieser Daten. Während dieser Vorgang keine Auswirkungen auf aggregierte Daten und Standardberichte hat, birgt er Herausforderungen für detaillierte Analysen. Insbesondere bei langfristigen Betrachtungen kann diese Begrenzung der Datenspeicherung zu Einschränkungen führen, die ärgerlich sein können. Insbesondere der Vergleich von bestimmten ereignisbasierten Kennzahlen im Jahresvergleich kann unter Umständen ohne BigQuery nicht erfolgen.

Sobald die Daten in BigQuery exportiert sind, entfällt die Beschränkung der Speicherdauer. Die Daten bleiben dauerhaft erhalten, vorausgesetzt, die entsprechenden Gebühren für die Google-Cloud werden beglichen. Somit bietet BigQuery eine Lösung für die langfristige Datensicherung und den dauerhaften Vergleich von Daten und damit auch die Möglichkeit der Beobachtung von Trends.

Vorteil zwei: kein Sampling

Bei großen Datenmengen durch viele Ereignisse oder Betrachtung längerer Zeiträume kann es in GA4 zum Sampling kommen (Abbildung 1). Dabei wird nur ein Teil der abgefragten Daten betrachtet und auf die Gesamtmenge hochgerechnet. Je nachdem wie groß die Stichprobe ist, kann das Ergebnis mehr oder weniger von der Realität abweichen. Bei einer geringen Stichprobengröße ist es nicht zu empfehlen, wichtige Entscheidungen auf Grundlage dieser Daten zu treffen.

DER AUTOR

Michael Janssen ist Webanalyst bei der Analytics-Agentur SISU digital. Er beschäftigt sich leidenschaftlich damit, wie man Webdaten erfassen und nutzen kann.

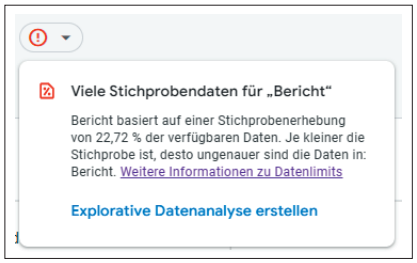

Abb. 1: Hinweis auf Sampling in GA4

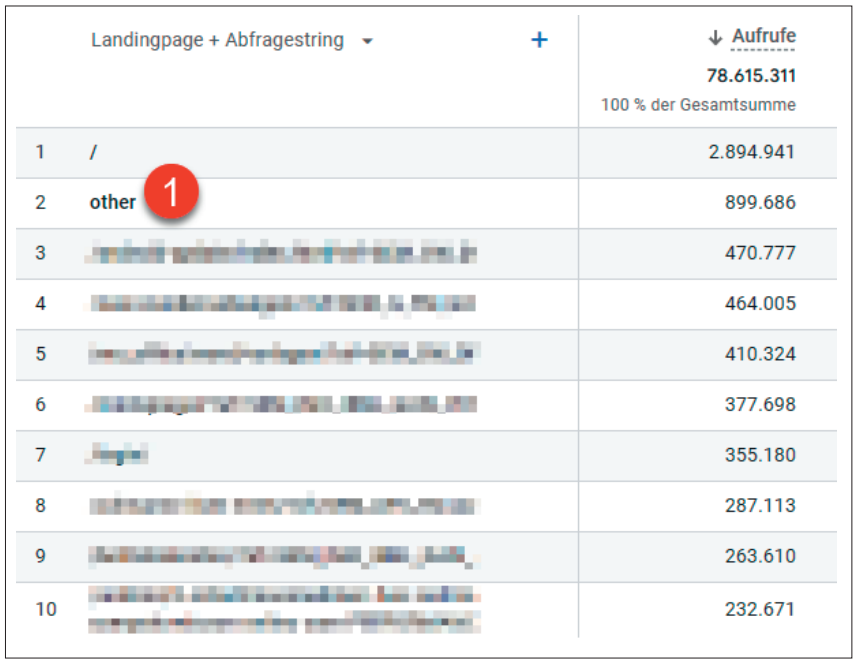

Abb. 2: Bericht mit der Zeile „other“

Bei den kostenlosen GA4-Propertys liegt das Limit für das Sampling bei zehn Millionen Ereignissen und für die GA4-360-Propertys bei 100 Millionen Ereignissen in den Berichten und bei einer Milliarde in der explorativen Datenanalyse. Auch wenn die Datenmengen sehr groß wirken, können diese Werte je nach Tracking-Konzept und Betrachtungszeitraum schnell erreicht werden.

Mit BigQuery gibt es das Problem des Samplings nicht, denn es werden immer sämtliche vorhandenen Daten abgefragt und in die Analyse einbezogen.

Vorteil drei: kein Problem mit zu vielen unterschiedlichen Dimensionswerten

Es ist immer wieder ärgerlich, wenn in einer Zeile „other“ auftaucht (Abbildung 2, Ziffer 1). Die Ausgabe dieses Werts als Dimensionswert hat als Grundlage eine zu große Vielfalt an unterschiedlichen Werten. Denn GA4 hat ein echtes Problem, wenn es zu viele unterschiedliche Dimensionswerte gibt.

In Universal Analytics war es in fortgeschrittenen Set-ups üblich, Dimensionen mit der Geräte-ID oder dem Zeitstempel zu erfassen. GA4 mag solche Dimensionen mit vielen unterschiedlichen Werten gar nicht gerne. Bei vielen unterschiedlichen Werten in einer einzelnen Dimension spricht man von hoher Kardinalität. Im Fall einer hohen Kardinalität fasst Google dann einen Teil der Werte in der Zeile „other“ zusammen.

Für Google beginnt eine Dimension, eine hohe Kardinalität zu haben, wenn es mehr als 500 unterschiedliche Werte pro Tag in einer Dimension gibt. Übrigens, ein ähnliches Problem gab es auch schon in Universal Analytics. Dort hatte es aber nur Auswirkungen, wenn diese Dimension im Bericht sichtbar war. In GA4 haben Dimensionen unter Umständen auch Einfluss, wenn sie im Bericht gar nicht sichtbar sind. Mit der Nutzung von BigQuery für die Datenauswertung kann man dieses Problem geschickt umgehen.

Vorteil vier: vielfältige Möglichkeiten bei der Detailanalyse

Mit den GA4-Daten können in BigQuery komplexe Abfragen und Analysen durchgeführt werden, die weit über das hinausgehen, was in der Standard-GA4-Oberfläche möglich ist. Zusätzlich können die Daten auch direkt mit anderen Quellen kombiniert und gemeinsam ausgewertet werden. Auch ist die Nutzung eigener Machine-Learning-Modelle möglich.

Vorteil fünf: Echtzeitauswertungen

Mit GA4 können die Daten in Echtzeit zu BigQuery exportiert werden. Das ermöglicht dann auch direkt eine Echtzeitauswertung, die gegebenenfalls für die Live-Optimierung von Seiteninhalten genutzt werden kann.

Obwohl der Einsatz von BigQuery für die Analyse von GA4-Daten viele Vorteile bietet, gibt es natürlich auch einige Nachteile, die beachtet werden müssen.

Nachteil eins: Unterschiede in den Daten

Ein wesentlicher Nachteil ist, dass die Daten in BigQuery immer einen Unterschied zu den Daten in der GA4-Oberfläche aufweisen werden. Das bedeutet, die Daten in BigQuery sind nahezu unverarbeitet. Algorithmen und Machine Learning verändern die Daten in der GA4-Oberfläche, aber nicht die Rohdaten. Das bedeutet, die Berechnung der Sessions und auch die Attributionslogik muss in BigQuery vorgenommen werden. Aber es gibt keinen Zugriff auf die intern in GA4 genutzten Modelle und Algorithmen. Ein gleichzeitiges Nutzen der Standardberichte in der Oberfläche und der Berechnungen

aus BigQuery kann also zu unterschiedlichen Werten und damit zu Kommunikations- und Entscheidungsproblemen führen.

Nachteil zwei: Kosten bei Speicherung und Analyse der Daten

Während die Funktion des Exports der Daten kostenlos ist, können aber beim Speichern und Analysieren Kosten entstehen. Für die ersten Schritte stellt Google ein kostenloses Kontingent bereit. Die ersten zehn Gigabyte fürs Speichern und das erste Terabyte für die Verarbeitung von Abfragen sind jeden Monat kostenlos. Darüber hinaus fallen Kosten an, die abhängig von der Menge der Daten sowie der Komplexität und dem Umfang der Abfragen sind.

Berechnung der Speicherkosten: Google gibt als Richtwert an, dass 600.000 Ereignisse circa ein Gigabyte an Speicherplatz belegen. Die Speicherung auf den Google-Servern in Frankfurt (Rechenzentrum: europe-west3) kostet pro Monat 0,023 US-Dollar. Dabei sind die ersten zehn Gigabyte pro Monat kostenlos und jede Tabelle, die 90 Tage nicht verändert wurde, wird nur mit den halben Speicherkosten in Rechnung gestellt. Unter diesen Voraussetzungen würde eine Seite mit 15.000 Ereignissen pro Tag in den ersten zwölf Monaten keine Kosten für die Speicherung erzeugen. Und selbst die Kosten für das gesamte zweite und dritte Jahr würden im einstelligen Dollar-Bereich bleiben.

Die Kostenberechnung für das Abfragen der Daten ist leider nicht so einfach möglich. Deshalb kann man direkt in der BigQuery-Oberfläche ausgeben lassen, wie viel Aufwand eine Abfrage erzeugen kann.

Neben den genannten Vor- und Nachteilen sind aber auch unbedingt folgende Punkte bei der Entscheidung und Nutzung von BigQuery zu beachten:

	Kleine Website (15.000 Ereignisse pro Tag)	Mittlere Website (100.000 Ereignisse pro Tag)	Große Website (500.000 Ereignisse pro Tag)
Erstes Jahr	0 US-Dollar	7,39 US-Dollar	36,67 US-Dollar
Zweites Jahr	3,14 US-Dollar	26,95 US-Dollar	131,4 US-Dollar
Drittes Jahr	8,24 US-Dollar	58,02 US-Dollar	283,83 US-Dollar

Abb. 3: Jährliche Kosten für die Speicherung der GA4-Daten in BigQuery

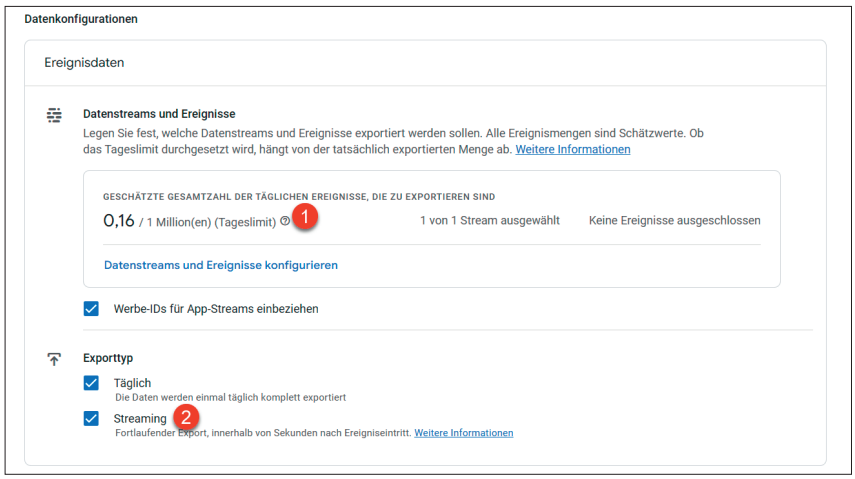

Abb. 4: Konfiguration des Datenexports

Benutzerdefinierte Abfrage eingeben

```
SELECT
  stream_id,
  PARSE_DATE("%Y%m%d", event_date) as date,
  EXTRACT(DATETIME FROM TIMESTAMP_MICROS(event_timestamp) AT TIME ZONE @timezone) as date_time,
  EXTRACT(HOUR FROM TIMESTAMP_MICROS(event_timestamp) AT TIME ZONE @timezone) as date_time_hour,
  EXTRACT(MINUTE FROM TIMESTAMP_MICROS(event_timestamp) AT TIME ZONE @timezone) as date_time_minute,
  FORMAT("%02d", DATETIME_DIFF(CURRENT_DATETIME(@timezone), EXTRACT(DATETIME FROM
TIMESTAMP_MICROS(event_timestamp) AT TIME ZONE @timezone), MINUTE)) AS minute_diff,  user_pseudo_id,
  (SELECT event_params.value.int_value FROM UNNEST(event_params) event_params WHERE event_params.key =
'ga_session_id') AS session_id,
  event_timestamp,
  event_name,
  CONCAT(user_pseudo_id, "_", event_timestamp, "_", event_name) AS unique_event_id,
  (SELECT event_params.value.string_value FROM UNNEST(event_params) event_params WHERE event_params.key =
'page_title') AS page_title,
  (SELECT event_params.value.string_value FROM UNNEST(event_params) event_params WHERE event_params.key =
'page_location') AS page_location,
  device.*,
  geo.*,
  traffic_source.*,
  ecommerce.*,
  items.*
FROM
  `xxx.analytics_123.events_*`,
UNNEST(items) items
WHERE SUBSTR(_TABLE_SUFFIX, -8) BETWEEN @DS_START_DATE AND @DS_END_DATE
```

Abb. 5: Abfrage, die bei unbedarfter Nutzung bei großen Projekten große Kosten verursachen kann

Daten erst ab Verknüpfung

Die Daten laufen erst in BigQuery ein, wenn die Verknüpfung erfolgt ist. Das ist besonders erwähnenswert, weil das bei GA Universal 360 anders war. Denn dort gab es den sogenannten Backfill. Dort wurden auch alte Daten eingespielt. Deshalb sollte frühzeitig überlegt werden, ob BigQuery für GA4 genutzt werden sollte. Insbesondere wenn die Kosten überschaubar sind, könnte ein vorsorglicher Export von Vorteil sein.

Nur ein Exportziel möglich

Jede GA4-Property kann nur mit einem BigQuery-Export verknüpft werden. Da nur ein Export zu BigQuery definiert werden kann, ist es wichtig, zu klären, wer der Besitzer der Daten sein soll. Oftmals nutzen Dienstleister aus Vereinfachungsgründen ihre eigenen Google-Cloud-Projekte als Ziel für die Speicherung. Dadurch sind diese aber Besitzer der Daten und beim Property-Inhaber landen keine Daten. Des-

halb sollte frühzeitig geklärt werden, was mit den Daten nach Beendigung der Zusammenarbeit geschieht und wie gegebenenfalls der Property-Inhaber die Daten in sein eigenes Google-Cloud-Projekt übertragen kann. Alternativ und auch sinnvoller ist direkt der Export in das Google-Cloud-Projekt des Property-Besitzers, auf das der Dienstleister Zugriff erhält.

Exportlimit bei einer Million Ereignisse pro Tag

Der kostenlose tägliche Export hat ein Limit von einer Million Ereignisse (Abbildung 4, Ziffer 1). Beim dauerhaften Überschreiten des Limits wird der Export nicht bei einer Million Ereignisse pro Tag gekappt, sondern der gesamte Export wird ab dem Folgetag pausiert. Erst wenn die Ereignisse wieder dauerhaft unter einer Million pro Tag sind, wird der Export fortgesetzt.

Es gibt zwei Möglichkeiten, wie dieses Limit umgangen werden kann. Zum einen mit dem Buchen von GA360 und zum anderen mit der Nutzung des Streaming-Exports (Abbildung 4, Ziffer 2). Beim Streaming-Export werden die Daten nahezu in Echtzeit in BigQuery gestreamt. Der Streaming-Export kostet aber pro Gigabyte an Daten aktuell zusätzlich 0,05 US-Dollar. Dafür stehen die Daten dann aber auch direkt für Echtzeit-Dashboards und -auswertungen zur Verfügung.

Vorsicht mit neuen Abfragen

Unbedachte oder schlecht optimierte Abfragen in BigQuery können schnell zu hohen Kosten führen (Abbildung 5). Es ist daher ratsam, sich mit den Grundlagen von SQL und den spezifischen Kostenstrukturen von BigQuery vertraut zu machen. Auch frei verfügbare oder kostenpflichtige LookerStudio-Templates können gegebenenfalls schnell hohe Kosten erzeugen. Bei großen GA4-Propertys mit vielen Daten in BigQuery kann ein Zugriff auf einen LookerStudio-Bericht auch mal mehrere Euro kosten. Das Gleiche gilt auch für den nativen BigQuery-Connector in LookerStudio. Deshalb unbedingt die Abfragen im Griff haben und sicherheitshalber die möglichen Kosten, die Nutzer erzeugen können, in der Google-Cloud begrenzen.

Fazit

Die GA4-Daten in BigQuery zu nutzen, bietet viele Vorteile und Möglichkeiten. Aber es ist wichtig, dass eine Nutzung von BigQuery mit Bedacht angegangen wird. Schnell können hohe Kosten erzeugt werden. Deshalb gilt gerade in Bezug auf BigQuery, dass ein solides Grundwissen einiges an Geld und Problemen sparen kann.