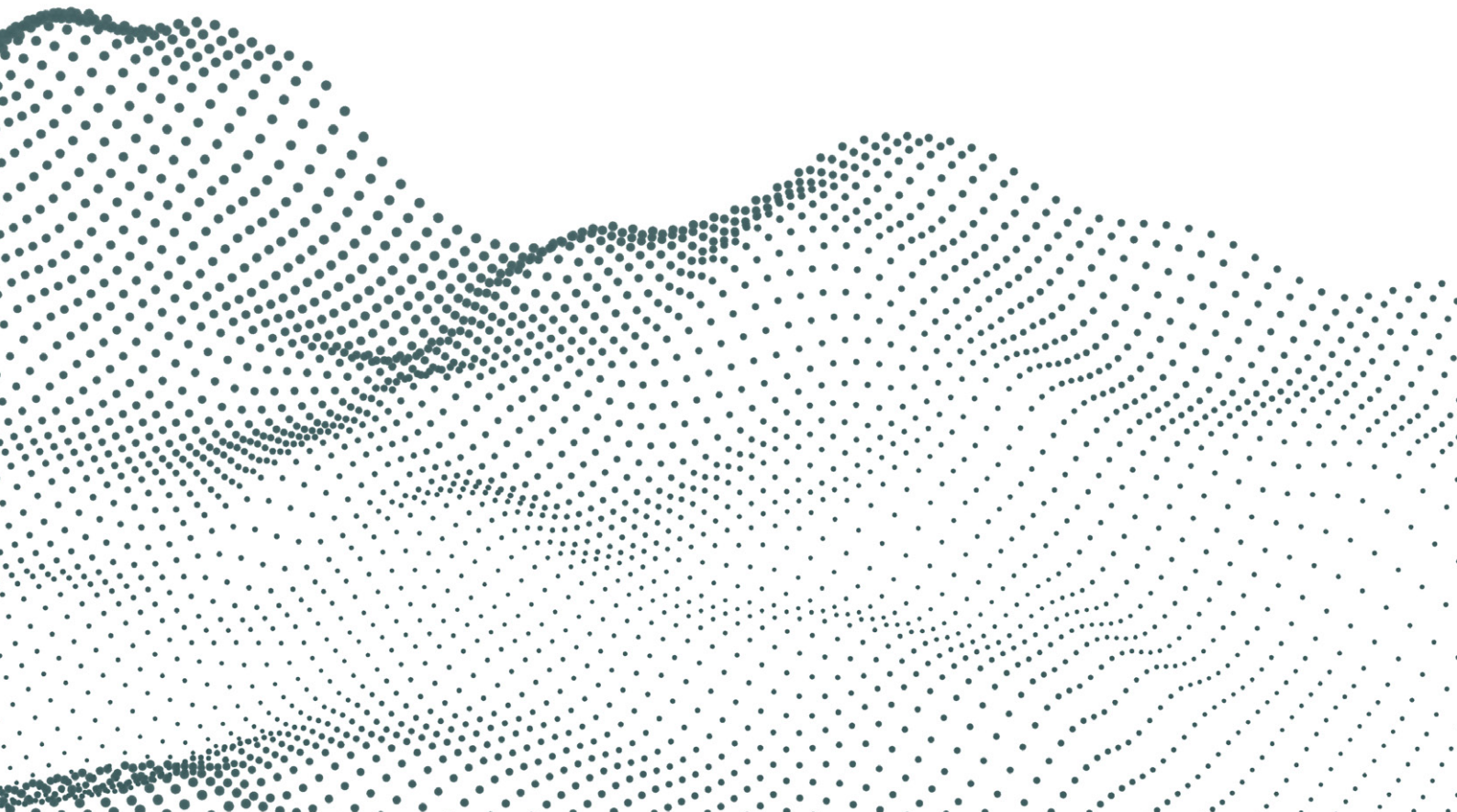
mohr stade

Rohdaten strukturieren mit Dataform

Vom Rohdatenchaos zu wertvollen Insights

Whitepaper



Inhalt

Veröffentlichung dieses Whitepapers: Februar 2025

1. Vorwort

Mit der Einführung von Google Analytics 4 (GA4) wurde der Zugriff auf Rohdaten für viele Nutzer erleichtert, insbesondere durch die direkte Integration in BigQuery. In diesem Zusammenhang bietet Dataform eine leistungsstarke Plattform, um Rohdaten zu transformieren und in nützliche Informationen umzuwandeln. Dieses Whitepaper untersucht die Funktionsweise und Vorteile von Dataform sowie seine Bedeutung für die effiziente Verwaltung moderner Datenpipelines.

2. Was sind Rohdaten und wo liegt ihr Potenzial?

Rohdaten stellen die unaufbereitete Form von Informationen dar, die direkt aus Quellsystemen stammen. Diese Daten sind oft denormalisiert, enthalten Fehler oder unvollständige Logiken und müssen transformiert werden, um in Berichten oder Analysen verwendet werden zu können.

Ein Beispiel für Rohdaten aus dem Analytics-Alltag sind GA4-Rohdaten. Diese werden von Google kontinuierlich in strukturierter Form an BigQuery übertragen und liegen dort in einem eventbasierten Format vor. Google nutzt dafür das Konzept der Sharded Tables: Alle Daten eines Tages werden in einer denormalisierten Tabelle gespeichert. Im Gegensatz zur Normalisierung, bei der Daten auf mehrere Tabellen verteilt werden, um Redundanzen zu reduzieren, werden bei denormalisierten Tabellen alle Informationen in einer einzigen Tabelle abgelegt. Jedes Event, wie z. B. ein Seitenaufruf oder ein Klick, wird dabei als eigenständige Zeile dargestellt. Um jedoch beispielsweise den zugehörigen Marketingkanal (Quelle) mit einem Key Event (Conversion) zu ermitteln, ist eine Transformation der Daten notwendig. Die Information für die Zuordnung des Key Events zu einem spezifischen Kanal ist in den Rohdaten nicht direkt enthalten. Im Vergleich zu herkömmlich aggregierten GA4-Daten bieten Rohdaten weitreichende Möglichkeiten zur Individualisierung und Erweiterung von Analysen:

Granularität und Flexibilität

Rohdaten enthalten alle einzelne Events und Parameter, wodurch tiefgehende Analysen auf Benutzer-, Sitzungs- oder Eventebene durchgeführt werden können. Während aggregierte Daten in GA4 stark durch vorab definierte Metriken und Dimensionen eingeschränkt sind, erlauben Rohdaten die Entwicklung eigener Logiken und die Kombination mehrerer Datenquellen.

Individuelle Attributionsmodelle

Mit Rohdaten können Unternehmen eigene Attributionsmodelle entwickeln und anwenden, die gut auf ihre spezifischen Marketingstrategien abgestimmt sind. Zum Beispiel wird die Modellierung einer benutzerdefinierten Customer Journey möglich, die nicht auf Standardlogiken wie Last-Click-Attribution beschränkt ist.

Fehlererkennung und Datenqualität

Rohdaten erlauben es, Ungenauigkeiten oder Anomalien früher zu identifizieren. So kann beispielsweise untersucht werden, ob die Daten bestimmter Marketingkanäle korrekt getrackt wurden oder ob unerwartete Werte wie fehlerhafte Umsätze in den Daten enthalten sind.

Kombination mit externen Datenquellen

Durch die Verbindung von Rohdaten mit anderen Plattformen wie CRM-Systemen, E-Commerce-Plattformen oder Drittanbieter-Tools lassen sich umfassendere Einblicke gewinnen. Diese Erweiterung ist mit aggregierten Daten in der Regel nicht möglich, da diese bereits stark reduziert vorliegen.

Langfristige Datenverfügbarkeit

In BigQuery gespeicherte Rohdaten können unbegrenzt archiviert werden, während in GA4 aggregierte Daten oft nur über einen eingeschränkten Zeitraum verfügbar sind. Die unbegrenzte Datenverfügbarkeit ermöglicht langfristige Analysen und den Aufbau historischer Modelle.

Zusammengefasst bieten Rohdaten durch ihre granulare Struktur und Flexibilität die Grundlage für individuelle, fundierte und verlässliche Analysen, die über die Möglichkeiten aggregierter Daten hinausgehen.

3. Rohdaten transformieren mit Dataform

Das Potenzial von Rohdaten zeigt sich besonders in der Möglichkeit, verschiedene Datenquellen miteinander zu kombinieren, anstatt sie isoliert zu betrachten. Ein Beispiel ist die Nutzung des Google-Ads-Rohdaten-Exports über die Data Transfer API in BigQuery, um die Rohdaten mit GA4-Daten zu verbinden. Auch die Integration weiterer Datenquellen, wie z. B. von Meta oder LinkedIn, kann wertvolle Einblicke bieten.

Tools wie Fivetran oder Supermetrics erleichtern die Prozesse durch unkomplizierte No-Code-Integrationen bei allen großen Public-Cloud-Anbietern. Mit einer solchen Datenverknüpfung ist es möglich, eine umfassende Sicht auf alle Marketingkanäle zu erhalten. Diese können dann anhand einheitlicher Logiken bewertet werden, ohne den Vorgaben einzelner Plattformen folgen zu müssen. Das Ergebnis kann beispielsweise ein Reporting sein, das alle relevanten Marketingkanäle inklusive Kostendaten einheitlich abbildet. Wie sich ein solches Vorhaben in der Google Cloud Platform effizient umsetzen lässt, wird nun anhand der Funktionalitäten von Dataform betrachtet.

3.1 Strukturierte Datenverarbeitung mit SQL-Pipelines

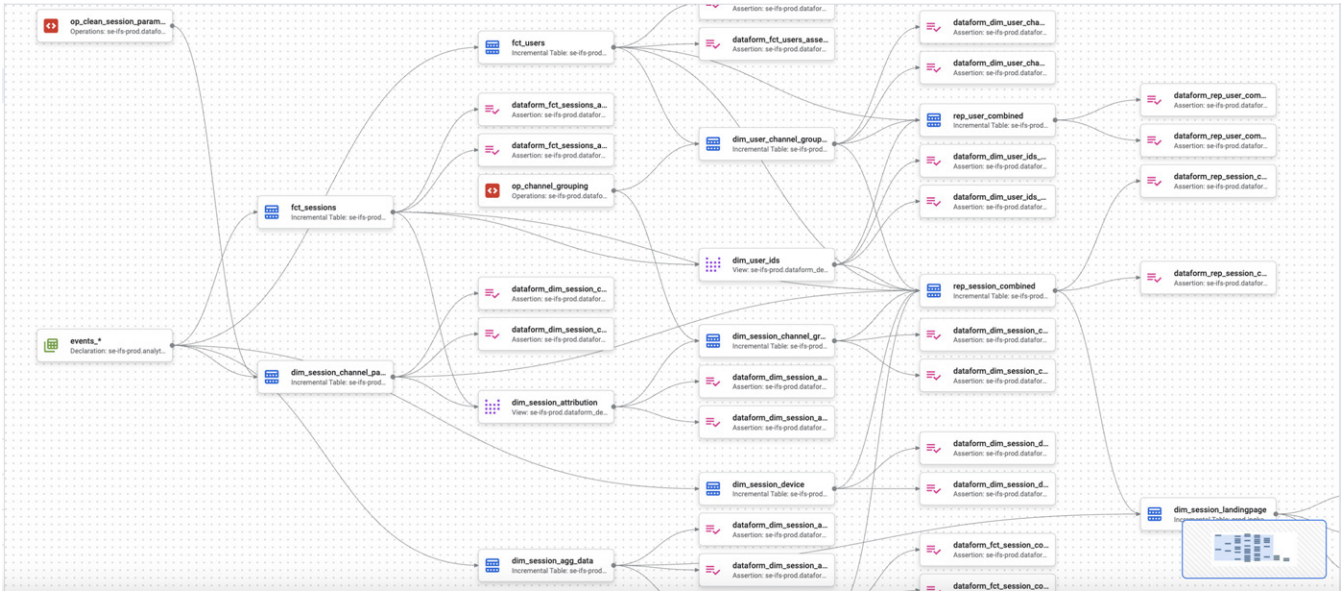


Abbildung 1: Compiled Graph einer Dataform SQL-Pipeline

Dataform bietet eine Umgebung für die effiziente Erstellung und Verwaltung SQL-basierter Datenpiplines. Entwickler und Analysten können hier Transformationen in einer modularen Weise erstellen, während Dataform die Abhängigkeiten zwischen diesen Transformationen automatisch erkennt und steuert. Dies wird durch die visuelle Darstellung im sogenannten "Compiled Graph" (Abbildung 1) erleichtert, der die Reihenfolge und Abhängigkeiten der Verarbeitungsschritte transparent macht.
Besonders hervorzuheben ist die Unterstützung inkrementeller Tabellen. Diese Funktion ermöglicht es, lediglich neue oder geänderte Daten zu verarbeiten, ohne die entsprechende Tabelle komplett neu aufzubauen. Dies wiederum spart Zeit und Rechenressourcen, was vor allem bei großen Datensätzen von Vorteil ist.

Der Vorteil liegt also in der Effizienz: Pipelines laufen schneller und können dadurch kostengünstiger und häufiger ausgeführt werden. Durch das Micro-Batching wird zudem die Verzögerung in nachgelagerten Prozessen deutlich reduziert, da nur die Änderungen seit dem letzten Lauf verarbeitet werden. Um sicherzustellen, dass keine doppelten Einträge in die Tabelle gelangen, wird häufig ein sog. *rowKey* eingesetzt. Dieser wird im config{} Block definiert und sorgt dafür, dass jede Zeile eindeutig identifiziert werden kann.

```
concat(DATE(extract(year from sessionDate), month, date_diff(sessionDate, firstSessionDate, day)+1),"-",browser, "-",deviceCategory) as rowKey
```

Abbildung 2: Beispiel-Definition eines rowKeys

Incremental Tables werden oft eingesetzt, um Tabellen mit Streaming-Daten oder sich regelmäßig ändernden Slowly Changing Dimensions (SCD) zu aktualisieren. SCDs ermöglichen das Festhalten des Zustands einer Dimension zu einem bestimmten Zeitpunkt mithilfe eines Snapshots. Dies erleichtert den Vergleich mit früheren Zuständen und erlaubt es, Datenveränderungen im Zeitverlauf nachzuvollziehen. Damit lassen sich beispielsweise der Statuswechsel eines Nutzers oder die historische Entwicklung des Lieferstatus einer Bestellung rückblickend analysieren.

3.2 Sicherstellung der Datenqualität durch automatisierte Tests

Ein zentrales Merkmal von Dataform ist die Integration von Tests, sogenannter "Assertions", die die Qualität der Daten in jedem Verarbeitungsstadium sicherstellen. Diese Tests können zum Beispiel prüfen, ob bestimmte Felder eindeutig sind. Also, ob einzelne Werte nicht leer sind oder ob spezifische Geschäftslogiken eingehalten werden. Diese Assertions sind flexibel und können an die individuellen Anforderungen eines Projekts angepasst werden. So kann unter anderem untersucht werden, ob Umsatzwerte plausibel sind oder ob alle Transaktionen mit mindestens einem Produkt verknüpft sind.

In Dataform werden Assertions in einem eigenen Verzeichnis (*/definitions*) gespeichert. Dort können verschiedene Arten von Assertions selbst definiert oder bereits vordefinierte Assertions wie *uniqueKey(s)*, *nonNull* oder *rowConditions* genutzt werden. Es folgt ein Beispiel zur Veranschaulichung:

```
assertions: {
uniqueKey:["session_id"],
nonNull:["session_id"],
rowConditions:[
"revenue > 0 and array_length(items)!= 0"
]
}
```

Abbildung 3: Beispiel-Code-Definitionen mehrerer Assertions, um sicherzugehen, dass die Revenue größer 0 und die Anzahl der Items ungleich 0 ist

In Abbildung 3 werden drei Assertions für die Transformation der Session erstellt. Dabei wird geprüft, ob die Spalte "session_id" eindeutig ist und keine leeren Werte enthält. Zudem wird das Feld "revenue" darauf getestet, ob es einen Bestellwert größer als Null aufweist. Abschließend wird kontrolliert, ob das Items-Array mindestens ein Produkt enthält. Aus dem im JavaScript-Code definierten config-Block generiert Dataform automatisch drei separate Tests. Diese werden nach der Ausführung der SQL-Abfrage eigenständig aufgeschlüsselt und ausgeführt. Für die Prüfung auf “Einzigartigkeit” der “session_id” generiert Dataform folgende SQL-Abfrage:

```
SELECT
*
FROM (
SELECT
session_id,
COUNT(1) AS index_row_count
FROM
`dataform-analytics-pioneers.dataform.stg_last_non_direct`
GROUP BY
session_id
) AS data
WHERE index_row_count > 1
```

Abbildung 4: Automatisch generierte SQL-Query einer Assertion, die auf Einzigartigkeit der "session_id" prüft

In manchen Fällen genügen die vordefinierten Assertions nicht, da spezifischere oder komplexere Prüfungen erforderlich sind. Dann können individuelle SQL-Abfragen erstellt werden, die als Tests definiert werden. Ein Beispiel wäre, nur die Zeilen des GA4-Datensatzes zu überprüfen, zu welchen Produktinformationen vorliegen sollten. Die nachfolgende SQL-Abfrage untersucht mehrere Aspekte des öffentlichen GA4-Datensatzes des Google Merchandise Stores:

```
select
*
from(
select
event_name
,count(1) as row_count
from
`bigquery-public-data.ga4_obfuscated_sample_ecommerce.events_*`
where
_table_suffix between '20210101' and '20210131'
and  event_name  in  ("view_promotion","view_item_list","view_item","select_item","add_to_cart","add_
to_wishlist","remove_from_cart","view_cart","begin_checkout","add_payment_info","add_shipping_
info","purchase")
and (
(
(array_length(items) = 0)
or EXISTS(SELECT * FROM UNNEST(items) AS items WHERE items.item_id is null)
or EXISTS(SELECT * FROM UNNEST(items) AS items WHERE items.item_name is null)
)
or (event_name = "purchase"
and (
ecommerce is null
or ecommerce.transaction_id is null
or ecommerce.purchase_revenue is null
or ecommerce.purchase_revenue <= 0
)
)
)
group by 1
)
where
row_count > 1
```

Abbildung 5: SQL-Query zur Prüfung von Standard-GA4-E-Commerce-Events

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

Row	event_name	row_count
1	purchase	300
2	add_payment_info	2841
3	add_shipping_info	3952
4	view_promotion	13385
5	view_item	26221
6	view_item_list	8

Abbildung 6: Ergebnisse aus der vorherigen SQL-Query

Das Ergebnis der SQL-Abfrage zeigt eine Liste der fehlerhaften oder unvollständigen GA4-Standard-E-Commerce-Events.
In der Praxis ist es sinnvoll, den Anteil der fehlerhaften E-Commerce-Events an der Gesamtzahl der E-Commerce-Events zu berechnen. Ohne diese Information ist es oft schwierig, die Bedeutung oder das Ausmaß des Fehlers richtig einzuschätzen.

JOB INFORMATION | RESULTS | JSON | CHART | PREVIEW

Row	event_name	bad_ecc_events	ecc_events
1	view_item	26221	86971
2	view_promotion	13385	53885
3	add_shipping_info	3952	3952
4	add_payment_info	2841	2841
5	purchase	300	1204
6	view_item_list	8	9
7	begin_checkout	0	11034
8	select_item	0	10229
9	add_to_cart	0	15522

Abbildung 7: Gegenüberstellung der gesamten ecc Events und der fehlerhaften ecc Events

In dem in Abbildung 7 gezeigten Beispiel wurden 300 fehlerhafte purchase Events festgestellt, was knapp 20 % aller purchase Events in diesem Zeitraum entspricht. Daher ist es sinnvoll, Schwellenwerte festzulegen, um zu bestimmen, ab welchem Fehleranteil ein kritischer Bereich erreicht wird.

3.3 Dataform Transformation Workflow im Cloud-Ökosystem

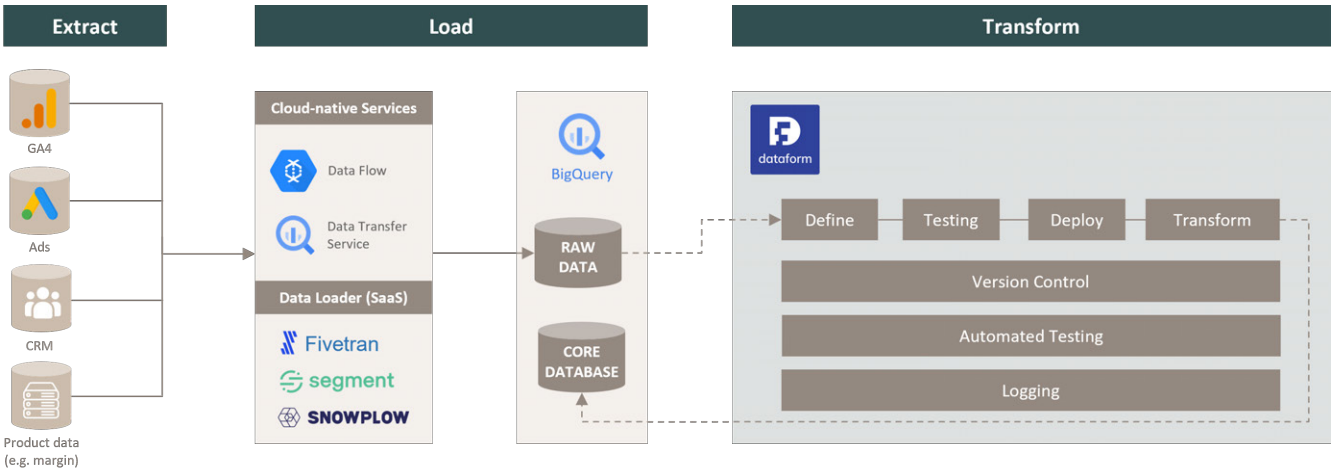


Abbildung 8: Dataform Transformation Workflow im Cloud-Ökosystem

Die Abbildung 8 veranschaulicht den technischen Workflow von Dataform innerhalb der Google Cloud. Daten werden mithilfe von Data Loadern wie Fivetran, Segment, Snowplow oder GCP-Diensten (z. B. Dataflow oder BigQuery Data Transfer Service) aus Quellsystemen nach BigQuery übertragen. Dort liegen sie im Rohformat als Tabellen oder Views vor. Ab diesem Punkt übernimmt Dataform die Rolle einer Orchestrierungs- und Modellierungsschicht. Die in Dataform erstellten Modelle (SQLX) verarbeiten die Rohdaten direkt im Data Warehouse, ohne dass die Transformation an der Quelle erfolgt. Nach der Verarbeitung und dem Testing werden die Daten in einer neuen Tabelle gespeichert, die für nachgelagerte Anwendungen wie Marketing- und Analysetools bereitgestellt wird.

4. Abgrenzung von Dataform zu Data Build Tool (dbt)

Seit Ende 2022 ist Dataform direkt in Google BigQuery integriert und somit ein zentraler Bestandteil der Google Cloud Platform (GCP). In Bezug auf die Funktionalität gibt es nur wenige Unterschiede zum Data Build Tool (dbt), da Dataform ebenfalls Git für Version Control sowie die Dokumentation von Modellen und Transformationen verwendet.

Ein klarer Vorteil von Dataform ist die reibungslose Integration in BigQuery, wodurch keine eigene Serverinfrastruktur notwendig ist. Im Gegensatz dazu erfordert dbt einen separaten Server. Dataform verwendet zudem eine JavaScript-API, die eine interaktive Arbeit mit SQL-Modellen ermöglicht, während dbt auf Jinja basiert, einer Templating-Sprache mit Python-ähnlicher Syntax. Ein Nachteil von Dataform ist die kleinere Community im Vergleich zu dbt, die eine geringere Verfügbarkeit vorgefertigter Packages mit sich bringt. Dennoch lassen sich viele dbt-Packages, wie etwa das GA4-Community-Package, mit leichten Anpassungen auch in Dataform verwenden.

Für Unternehmen, die bereits mit GCP und BigQuery arbeiten, kann Dataform langfristig eine bessere Integration bieten. Die Entscheidung zwischen Dataform und dbt sollte allerdings immer von den spezifischen Unternehmensanforderungen und den vorhandenen Kompetenzen im Data-Team abhängen.

5. Unsere Empfehlung

Die Arbeit mit Rohdaten bietet viele Vorteile. Sie ermöglicht die Kombination verschiedener Datenquellen, die Erstellung eigener Logiken und bietet maximale Flexibilität im Umgang mit den Daten. Außerdem behalten Unternehmen die volle Kontrolle und können ihre Daten bei Bedarf problemlos in andere Dienste migrieren. Dadurch sind sie nicht dauerhaft an eine Public Cloud oder einen Anbieter gebunden. Tools wie Dataform erleichtern die technische Umsetzung erheblich, sodass der Fokus auf die Erstellung von SQL-Abfragen gelegt werden kann. Dank der nahtlosen Integration in BigQuery und die Google Cloud Platform ist besonders im Bereich Marketing Analytics ein einfacher Einstieg möglich. Rohdaten aus Google Analytics 4 oder Google Ads können ohne den Einsatz von Drittanbietertools kostenlos nach BigQuery exportiert werden. Unternehmen, die ohnehin viele Google-Tools nutzen, profitieren von der Nutzung von Dataform in der Google-Cloud-Umgebung. Es ist jedoch wichtig zu bedenken, dass die Arbeit mit Rohdaten auch Herausforderungen mit sich bringt. Die große Flexibilität bedeutet gleichzeitig eine hohe Eigenverantwortung. Die Qualität der Transformationen muss selbständig überprüft werden. Außerdem kann die Implementierung neuer Funktionen sehr zeitintensiv sein. Trotz dieser Herausforderungen lohnt sich die Investition in Rohdaten in vielen Fällen. Mit dem wachsenden Fokus auf Data Ownership, erhöhter Verantwortung und den steigenden Anforderungen Analyse Funktionen bietet die Nutzung von Rohdaten Unternehmen langfristig einen klaren Vorteil.

Über mohrstade

Unternehmen

mohrstade ist eine Beratung für Marketing-Technologie in München, Hamburg und Wien. mohrstade ist spezialisiert auf Projekte in den Bereichen Data Collection, Data Management, Analytics, Marketing Activation und Data Visualization. Diese Services bietet mohrstade in zertifizierten Partnerschaften mit Marketing-Software-Herstellern an.

Managing Partner

Patrick Mohr

Co-Founder & Managing Partner

Patrick ist Gründer und Geschäftführer von mohrstade. Bereits während seines Studiums für BWL, Finance und Information (MSc) sammelte er Erfahrungen im Management Consulting. Später arbeitet er als SEA Manager, Data Scientist und Analytics Consultant bei Rocket Internet, Group M und UDG. 2017 baute er schließlich den Münchner Standort von Trakken auf. Parallel arbeitet er als Dozent an Universitäten. Darüber hinaus ist er Co-Organisator von Analytics Pioneers, der größten Analytics Community im DACH-Raum.

patrick@mohrstade.de

Marcus Stade

Co-Founder & Head of Analytics

Marcus ist Gründer von mohrstade und Head of Analytics. Darüber hinaus ist er Co-Organisator von Analytics Pioneers, der größten Analytics Community im DACH-Raum. Zuvor hat er im Bereich Web-Development und Online-Marketing gearbeitet. Auf seinem Blog www.marcusstade.de schreibt er regelmäßig zu Themen der Digitalen Analyse.

marcus@mohrstade.de

mohr
stade

Mohr & Stade GmbH
Friedrichstraße 1A
80801 München

www.mohrstade.de