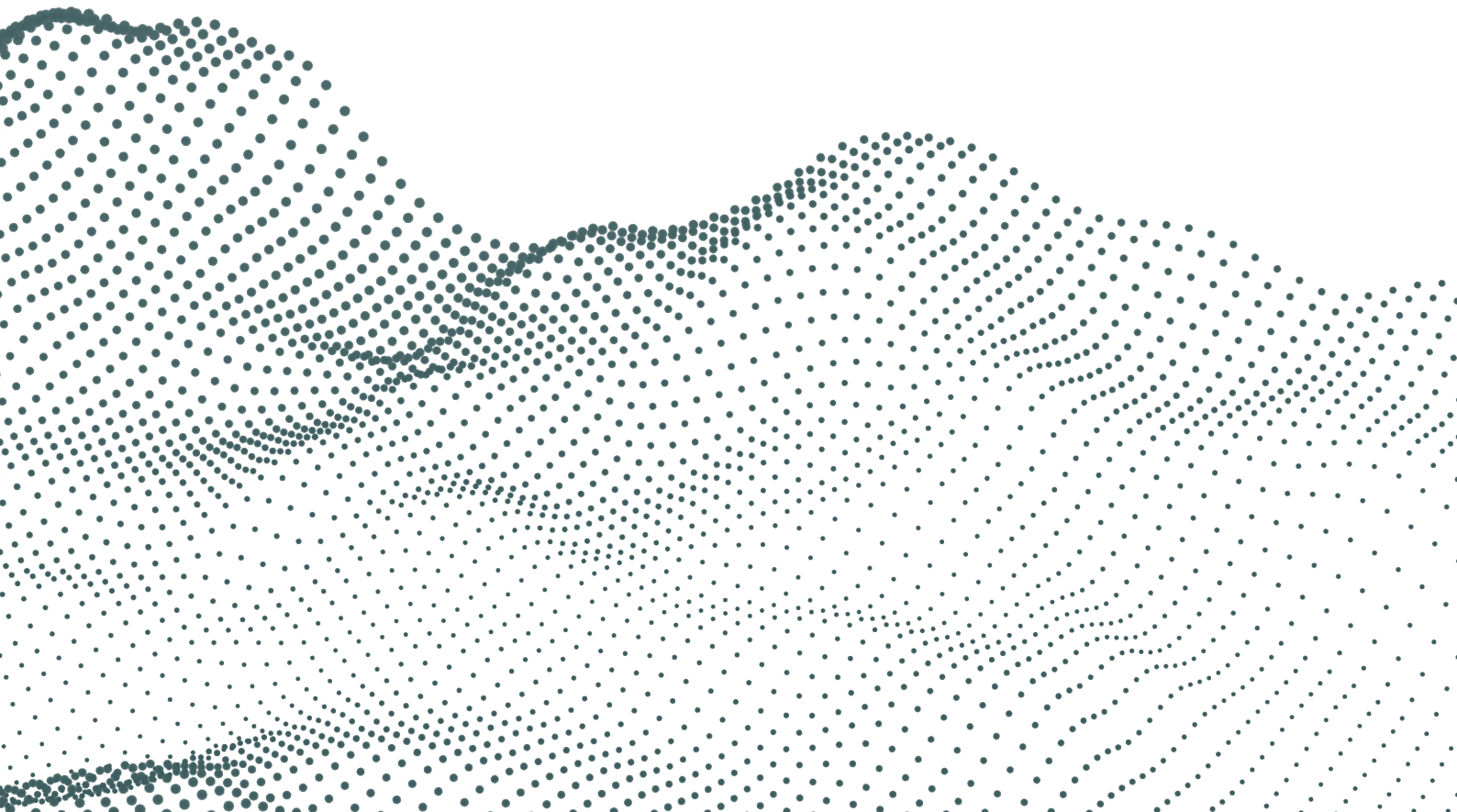
mohr stade

Nutzung von Matomo-Daten in der Google Cloud Platform

Import, Anreicherung und Nutzung von Matomo-Daten in Google BigQuery

Whitepaper



Inhalt

1. Einleitung 3
2. Über Matomo 4
   2.1 Besonderheiten des Tools 4
   2.2 Einschränkungen und Grenzen 5
3. Nutzung der Matomo-Daten in der Google Cloud Platform (GCP) 7
4. Empfehlung 8

Über mohrstade 9

Veröffentlichung dieses Whitepapers: Juni 2022

1. Einleitung

Datenschutzverordnungen wie DSGVO, ePrivacy und das TTDSG üben zunehmenden Druck auf die Datenerfassung durch Website-Betreiber und Werbetreibende aus. Viele Unternehmen stellen ihre Analytics- und Marketing-Tool-Landschaft deshalb auf die Probe und evaluieren Alternativen. Matomo bietet ein Self- und EU-Hosting des Servers, womit das Tool definitiv eine gute Lösung für die Einhaltung gegebener Datenschutzrichtlinien darstellt. Jedoch stößt das Hosting der Daten auf einem eigenen MySQL-Server mittel- bis langfristig an seine Grenzen. Für komplexere Analysen und Daten-Modellierung ist die Datenbank nicht ausgelegt. Die Auslagerung der Daten in eine skalierfähige Cloud-Umgebung ist demnach ein notwendiger Schritt, wenn tiefgreifende Erkenntnisse über den Nutzer generiert und crossfunktional mit den Daten gearbeitet werden soll. In diesem Zusammenhang bieten Cloud-Umgebungen wie z.B. die Google Cloud Platform mit Google BigQuery die Nutzung einer skalierbaren Daten-Infrastruktur aus Rechenleistung, Datenspeicherung und komplementären Services nach dem “Pay per Use”-Modell.

Dieses Whitepaper soll daher die Möglichkeiten für die weitere Datenverarbeitung und Analyse der Matomo-Daten innerhalb der Google Cloud Platform aufzeigen, um die Grenzen einer reinen Matomo Interface-Nutzung zu überschreiten.

2. Über Matomo

Matomo ist ein Open-Source Analytics Tool (Library) und wurde bis Anfang 2018 noch unter dem Namen Piwik geführt. Der wohl größte Unterschied zu anderen Analytics Tools wie bspw. Google Analytics ist das Self-Hosting erhobener Daten. Die Integration und Implementierung des Trackings ist etwas aufwendiger, bietet jedoch gleichzeitig volle Kontrolle über den eigenen Datenbestand. So kann bei der Speicherung sensibler Daten mehr Privatsphäre für den Nutzer gewährleisten werden, da die Daten nicht automatisch mit den Server-Endpoints von Drittanbietern geteilt werden. Mit dem Fokus auf Privatsphäre und Datenhoheit bedingt das Tool in anderen Bereichen allerdings auch Einschränkungen. Im folgenden Kapitel werden grundlegende Besonderheiten und Grenzen von Matomo vorgestellt.

2.1 Besonderheiten des Tools

On-Premise und Datenschutz

Das Self-Hosting der Daten auf einem eigenen Server bietet vollständige Kontrolle über die Erhebung und Weiterverarbeitung der Daten. Im ersten Schritt wird sichergestellt, dass erhobenen Daten nicht mit Drittanbietern geteilt werden. Dies stärkt nicht nur den First Party-Datenbestand, sondern erleichtert auch die Gestaltung der Datenschutzerklärung nach der DSGVO. Zudem besteht die Möglichkeit, über das User-Interface ein sogenanntes “Cookieless-Tracking” zu aktivieren. Die Identifizierung des Nutzers erfolgt anhand der von Matomo vergebenen *config_id*, welche einzelne Interaktionen innerhalb eines 24-Stunden-Zeitfensters einem Website-Besucher zuordnet. Der von der *config_id* übergebene String basiert auf einem gehashtem Fingerprint des Browsers - bestehend aus der anonymisierten IP-Adresse, dem Betriebssystem, individuellen Spracheinstellungen und verwendeten Browser-Plugins -, welcher laut Matomo das Tool von seiner gesetzlichen Opt-In-Pflicht entbindet.

Open Source

Eine Besonderheit gegenüber anderen Analytics-Lösungen besteht darin, dass es sich bei Matomo um eine Open Source-Library handelt. Der Quelltext des Tools ist somit öffentlich einsehbar. Erweiterungen und Plug-ins werden von der Community selbst entwickelt und über den Matomo Marketplace von Matomo und seinen Entwicklern angeboten. Zu den Funktionen und Features, die von Matomo standardmäßig gestellt werden, kommen auf diese Weise viele Zusatzfunktionen hinzu, mit welchen sich das Tool individuell und an die Bedürfnisse des eigenen Unternehmens und der Website anpassen lässt.

2.2 Einschränkungen und Grenzen

Cookieless-Tracking und das TTDSG

Seit der Einführung des TTDSG im Dezember letzten Jahres unterliegt auch das Fingerprinting als Methode für das “Cookieless-Tracking” strengeren Richtlinien. Hierbei kommt es auf Art und Umfang der Identifier an und ob für deren Erhalt lt. §25 TTDSG ein Zugriff auf die Endeinrichtung (Endgerät) erforderlich ist. Es gibt zurzeit jedoch noch viel Spielraum für Interpretation bei der Einstufung von Daten, die explizit abgefragt werden müssen. Dazu gehören u. a. individuelle Browser-Plug-ins, welche von Matomo beispielsweise abgefragt werden. Die Rechtslage ist häufig unklar und auf den Individualfall abzugrenzen. Auch sollte die Qualität der Identifier kritisch hinterfragt werden. Goldstandard stellt an dieser Stelle immer noch die Datenerhebung über First Party Cookie-Identifier dar. Es ist zu empfehlen, langfristig mehr Wert auf eine hohe Einwilligungs-Quote zu legen, um eine hohe Datenqualität zu gewährleisten.

Ein konkretes Vorgehen ist immer auf den Einzelfall auszulegen und sollte im Rahmen einer individuellen Rechtsberatung definiert werden.

Proof of Concept und Community Support

Tools wie Google Analytics haben weltweit mehrere hunderttausend Nutzer. Für die meisten Probleme und Herausforderungen bestehen mittlerweile gut dokumentierte Lösungsansätze mit dem Support einer großen Community. Am Markt gibt es für Google Analytics also wesentlich mehr Tutorials und Kurse für Einsteiger, um sich im Tool zurechtzufinden, als für Matomo. Des Weiteren profitieren Tools wie Google Analytics von Daten aus dem Google Display Netzwerk und anderen Diensten wie bspw. Google Ads. Hinzu kommen zahlreiche Kooperationen mit anderen Software-Herstellern, die das Produktportfolio ergänzen und ein breites Spektrum an kostenlosen Features bieten, die entweder in Matomo nicht existieren oder nur kostenpflichtig über den Matomo Marketplace erhältlich sind

Einrichtung und Wartung

Die Inbetriebnahme von Matomo erfordert die Installation des Tools auf einem eigenen Server. Das macht es gerade für Anfänger oder Benutzer ohne IT-Kenntnisse zunächst etwas schwierig, das System zu implementieren und zum Laufen zu bringen. Außerdem führt das Hosting auf eigenen Servern zwangsweise zu einem höheren Installations- und Set-up-Aufwand sowie zu höheren Folgekosten bei Wartung, Back-up und allgemeinen Serverkosten. Änderungen am Set-up sind also größtenteils mit direkten Änderungen an der technischen Implementierung verbunden.

Usability und Performance

Bekannte Tools wie Google Analytics bieten eine Vielzahl kostenloser Funktionen und Reports, die nur in der kostenpflichtigen Matomo-Cloud nutzbar sind. Dazu zählen unter anderem das Erstellen von Custom Reports, der bekannte User-Behavior Flow sowie der Kohorten-Report, welche nur über Premium-Funktionen hinzugebucht werden können. Auch die Verknüpfung mit Marketing-Kanälen wie z. B. Google Ads ist im Vergleich zu Google Analytics bei Matomo aufwendiger. Wer an das Google Analytics-Interface gewöhnt ist, der wird bei Matomo einiges an Detailtiefe und Funktion in den Standard ABC-Reportings vermissen.

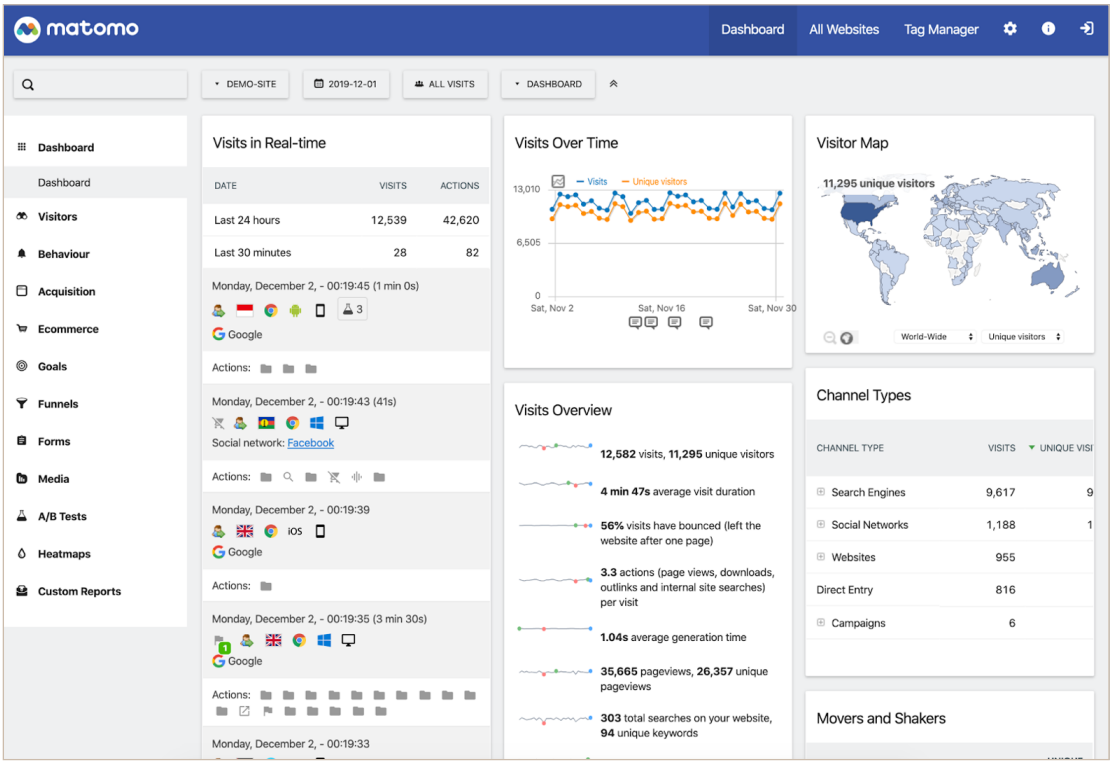


Abbildung 1: Matomo Interface - Overall Dashboard

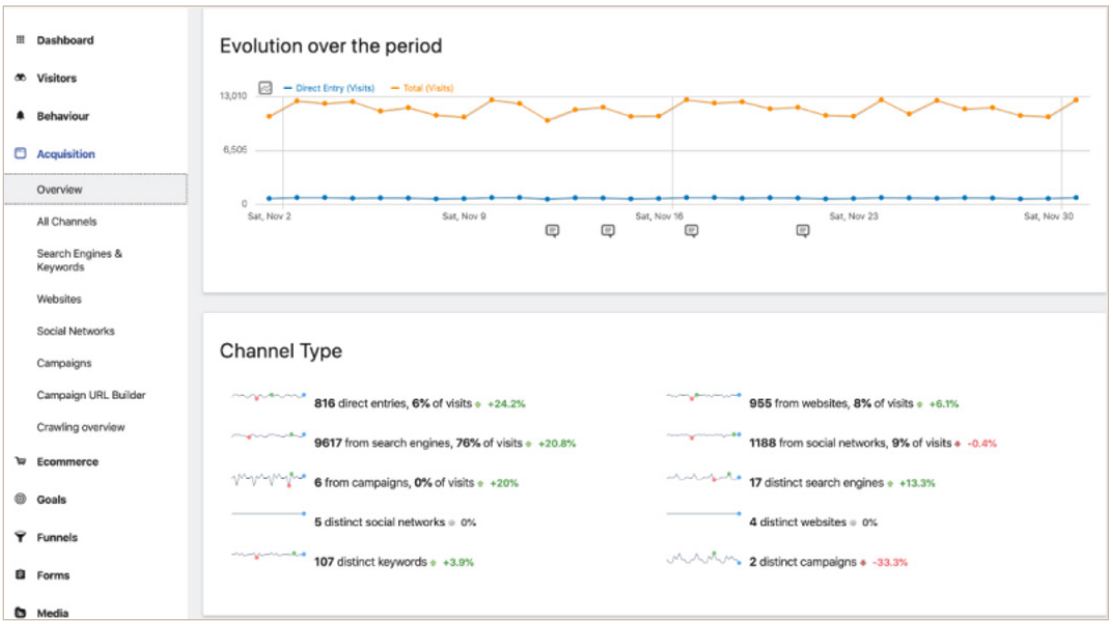
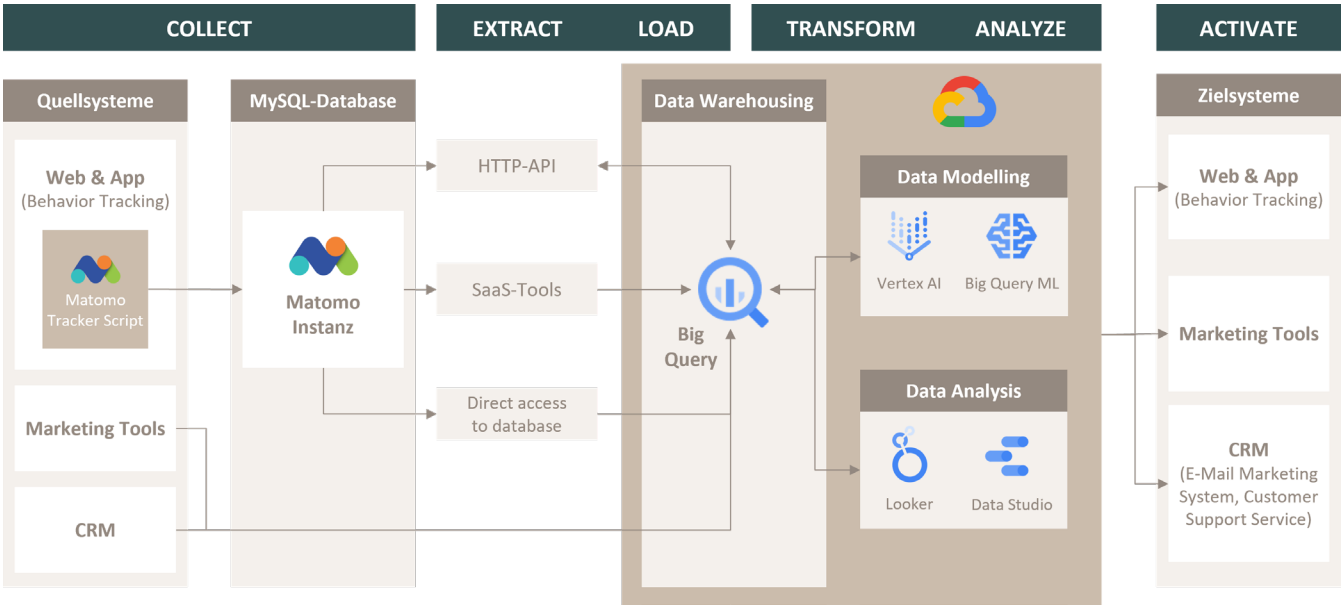


Abbildung 2: Matomo Interface - Acquisition Overview

Matomo kann als Stand-alone-Lösung allerdings eine ganze Menge leisten. Insbesondere wenn der Traffic auf der Webseite überschaubar ist und kein Bedarf an umfangreichen Segmentierungs- oder Analysefunktionen und Reportings gegeben ist, kann das Tool mit einem benutzerfreundlichen User-Interface glänzen. Gerade für Blogger oder einfache Firmenwebseiten ist Matomo häufig das Analytics-Tool der Wahl, um Besucher datenschutzkonform zu erfassen. Ist jedoch viel Traffic auf der Website und ein Anspruch an komplexe Analysen und Reportings gegeben, kann das Tool schnell an seine Grenzen geraten. Es gilt also die Vorteile, die Matomo mit sich bringt, durch eine skalierfähige Rechenleistung sowie Analyse- und Segmentierungsfunktionen zu erweitern. In diesem Zusammenhang bietet sich der Import der Matomo-Daten in ein Cloud-Data Warehouse wie Google BigQuery an, um auf ein breites Spektrum an Funktionen und Performance der Google Cloud-Umgebung zurückgreifen zu können. Im folgenden Kapitel wird ein Lösungsansatz für den Import der Matomo-Daten sowie ihre anschließende Weiterverarbeitung und Ergänzung in BigQuery vorgestellt.

3. Nutzung der Matomo-Daten in der Google Cloud Platform (GCP)

Zur Darstellung der Daten-Pipeline wird ein Architektur-Plot entwickelt, welches den Prozess und die Möglichkeiten für die Weiterverarbeitung der Matomo-Daten innerhalb der Google Cloud Platform (GCP) aufzeigt. Übergeordnet ist der Prozess in die Bereiche *Collect - Extract - Load* (Erheben, Extrahieren und Laden der Daten), *Transform - Analyze* (Modellierung und Analyse) und *Activate* (Nutzung der Daten) unterteilt, welche nachfolgend im Detail beschrieben werden:



COLLECT - EXTRACT - LOAD

Über das auf der Website oder in der App implementierte Matomo Tracker Script werden die Daten erfasst und an die ausgewählte Matomo-Instanz auf dem eigenen MySQL Server gesendet. Nun gibt es unterschiedliche Mittel und Wege, die Matomo-Daten aus der MySQL-Datenbank in ein Data Warehouse (z. B. BigQuery) zu laden. Die offiziellen Vorgehensweisen setzen jedoch technisches Know-how voraus. Standardmäßig bietet Matomo zwei Lösungen für das Extrahieren und Laden der Daten an.

1. HTTP-API

Mit der von Matomo zur Verfügung gestellten HTTP-API lassen sich alle Nutzerdaten und Nutzer-Interaktionen (Visits, Seiten, Klicks, Ereignisse, Downloads, Custom Dimensionen etc.) an jeden beliebigen HTTP-Endpoint exportieren. Auf die API kann unter Live.getLastVisitsDetails zugegriffen werden. Die API ermöglicht den Export genannter Nutzer- sowie Clickstream-Daten für eine bestimmte Website und ein bestimmtes Datum. Es folgt ein Beispiel für einen API-Call:

https://mohrstade.de/matomo/?module=API&method=Live.getLastVisitsDetails&idSite=62&period=day&date=yesterday&format=xml&token_auth=12345.6789&filter_limit=100

Für die Nutzung der API muss für jeden Matomo-Nutzer ein neuer Token erstellt werden, welcher für die Abfrage im Parameter &token_auth= abgebildet wird. Matomo selbst empfiehlt, einen neuen Nutzer mit "Read only"-Rechten für die API-Abfrage zu erstellen. Die Daten können in verschiedenen Formaten exportiert werden, darunter in JSON, XML und CSV. Zur Bestimmung des Formats des Exports muss lediglich der Parameter &formal= um das gewünschte Format angepasst werden. Der abgefragte Datensatz enthält bis zu 500 Interaktionen für jeden Visit. Sollte vermehrt die Anzahl der Visits über das Hit-Limit geraten, lässt sich die API anpassen.

2. Direktzugriff auf die MySQL Datenbank	Eine andere Lösung ist der direkte Zugriff auf die MySQL-Datenbank von Matomo Analytics (früher Piwik Analytics), in der alle Daten der Instanz gespeichert sind. Diese Lösung stellt eine performantere Option dar, um größere Datenmengen zu importieren, insbesondere wenn das Data Warehouse den Datenimport aus MySQL unterstützt (die meisten Data Warehouses tun dies, einschließlich BigQuery). Auch hier empfiehlt Matomo, einen MySQL-Nutzer mit “Read only”-Rechten anzulegen, der nur auf die Matomo-Datenbank zugreifen und Daten auslesen kann.
3. SaaS-Tools und Konnektoren	Anbieter wie bspw. Pipes bieten vorgefertigte Konnektoren, durch die sich über ein benutzerfreundliches Interface die Daten aus der Matomo Instanz zu BigQuery übermitteln lassen. Allerdings sollte beachtet werden, dass die meisten Saas-Tools nicht auf Matomo begrenzt sind, sondern ein breites Spektrum an Konnektoren anbieten. Dies hat zur Folge, dass das Pricing sich für den einzelnen Anwendungsfall nicht rentiert. Hinzu kommt, dass die Konnektoren nur auf die von Matomo gestellte HTTP-API zugreifen. Daher macht es eventuell Sinn, sich mit der technischen Dokumentation der API zu befassen, um mittelfristig einen eigenen Konnektor zu bauen.

TRANSFORM - ANALYZE

Innerhalb von BigQuery können die Daten anschließend mit ausgewählten Cloud-Services nach dem “Pay per Use”-Modell weiterverarbeitet und analysiert werden. In diesem Punkt liegt der größte Vorteil für die Auslagerung der Matomo-Daten in die Google Cloud. Mit Services wie der Vertex AI oder BigQuery ML lassen sich durch maschinelles Lernen komplexe Analysen und Segmentierung auf Basis der Nutzerdaten durchführen, um diese für das Dashboarding in einem performanten BI-Tool wie Looker oder Power BI sowie für die Marketing-Aktivierung in Tools wie Google Ads oder Facebook zu nutzen.

ACTIVATE

Die einheitlichen Nutzer-Segmente lassen sich anschließend mit Hilfe von build-in Konnektoren oder des BigQuery Data Transfer Services an die gewünschten Zielsysteme ausliefern und für bspw. Remarketing-Kampagnen verwenden.

4. Empfehlung

Es lässt sich festhalten, dass Matomo eine gute Lösung für ein DSGVO-konformes Web Tracking darstellt und somit Potenzial bietet, die bestehende Datenerfassung als Tool zu ersetzen. Jedoch fehlen dem Tool grundlegenden Analyse-Funktionen sowie die Möglichkeit, komplexere Analyse- und Segmentierungsfunktionen zu nutzen. Die Rechenressourcen der eigenen Datenbank kommen bei hoher Auslastung zusätzlich schnell an ihre Grenzen und sind relativ starr, wenn es um die Skalierfähigkeit der Rechenleistung geht. Die Basis der Datenverarbeitung und Analyse sollte daher eine performante Cloud-Umgebung wie die Google Cloud darstellen. Hierfür eignet sich Google BigQuery als Data Warehouse im Zentrum der GCP. Diese bietet nach dem “Pay per Use”-Modell eine Vielzahl an komplementären Services für die Durchführung komplexer Analysen und Segmentierungen auf Basis statistischer Verfahren oder maschinellen Lernens. Zusätzlich stehen eine Vielzahl an build-in Konnektoren und API-Schnittstellen zur Verfügung, um die Daten anschließend in ein performantes BI-Tool wie bspw. Looker oder Power BI zu überführen

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

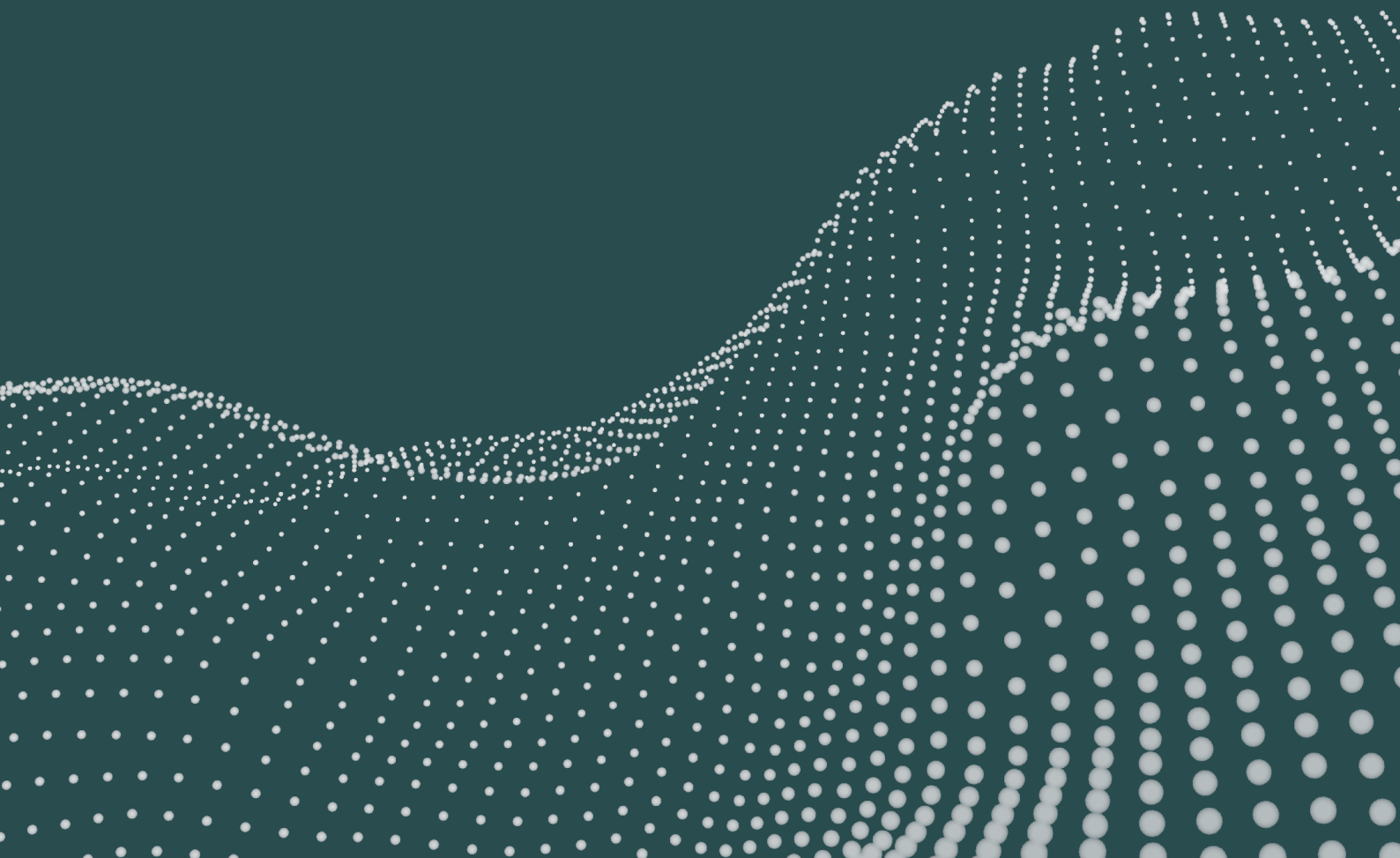
Marcus Stade

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
Landwehrstraße 2
80336 München

www.mohrstade.de