e·capacity
Digital Strategy and Analytics

Sådan trækker du rådata ud af Adobe Analytics, så de kan bruges til ML- og AI-analyse

Whitepaper, eCapacity, 2019. Af Troels Moltsen.

eCapacity, Sjæleboderne 4, 3. sal, 1122 København K, (+45) 70 26 27 23, info@eCapacity.dk

Indholdsfortegnelse

Forord

Der går næppe en dag, uden at nyhedsstrømmen flyder over begreber som Machine Learning, AI og predictive analytics. Med omtaler af metoder, teknologier og processer, der skal sikre, at data udnyttes, bindes sammen og gives forudsigelseskraft. Med gode råd og velmenende anbefalinger.

Men ofte bliver snakken væsentligt mindre klar, når det kommer til spørgsmålet om, hvad man konkret skal gøre for at udnytte de data, man som virksomhed har til rådighed. Ja, ofte får man indtrykket af, at antallet af redskabsråd, der i praksis kan hjælpe dig fremad i din dataudnyttelse, er omvendt proportionalt med den generelle buzz-word volumen.

Det har vi hos eCapacity besluttet os for at gøre noget ved. Vi arbejder til dagligt med vores kunders måske største datakilde: vores kunders websites og deres øvrige digitale kanaler. Vi implementerer enterprise analytics og tag managementsystemer, vi binder datakilder sammen med DMP'er og CDP'er, vi optimerer og laver avancerede data-analyser. Og vi hjælper med at få skabt overblik over strategiske mål, og hvad der skal gøres, for at de nås. Derfor ved vi, hvor skoen i virkeligheden trykker, når det gælder spørgsmålet om at få de mange bits og bytes til at blive til rigtige indsigter, at hive data ud af siloer og køre de rigtige algoritmer, der får dem til at tale sammen og give rigtig forretningsmæssig værdi.

I dette whitepaper deler vores Troels Moltsen helt lavpraktisk ud af vores viden. Hvis dit website kører Adobe Analytics, fortæller Troels her i detaljer, hvordan du får data eksporteret fra Adobe Analytics i en form, der gør dem velegnet til efterfølgende at køre Machine Learning og AI-algoritmer på dem. Whitepaperet er skrevet i udviklersprog, men skulle gerne give et overblik også til alle de af os, der ikke mestrer R eler Python by heart.

God læselyst. Og held og lykke med at skabe dataresultater i praksis.

København, den 28. august 2019

Andreas Petersson,
Partner og Director, Analyse og optimering

Introduktion

Her er situationen: Du vil gerne køre maskinlærings-algoritmer på dine Adobe Analytics-data, og måske endda kombinere disse data med data fra andre kilder. Men de data, Adobe stiller til rådighed for dig, er slet ikke finkornede nok til at gøre tricket. Eller det vil sige: det er de ikke pr. default. Men med lidt Adobe-massage kan der gøres meget. Bare følg guiden i dette whitepaper.

Forleden stod en af vores kunder med et problem. Page load - hastigheden siderne loadede på hans website - var uacceptabel høj. Han var bange for, hvordan det påvirkede besøgende, deres brugerrejser og i sidste ende hans salg.

For at hjælpe ham med at finde ud af, om der var nogen påvirkning, og hvad der i givet fald kunne gøres ved det, kiggede jeg på hans Adobe Analytics-installation, der lagrede al aktivitet fra alle brugere på hans site. Disse data ville helt sikkert være i stand til at fortælle os, hvilke brugere der faktisk oplevede det forsinkede pageload, og også hvordan pageloadet påvirkede konverteringen af de enkelte brugere. I særdeleshed ønskede vi at bruge ML og AI til at analysere den enorme mængde data, som Adobe-systemet rummer (jeg havde øjnene på open source XGBoost til dette job).

Adobe-data er ikke finkornet nok (i det mindste ikke pr. default)

For at klare tricks som disse skal du have data for at være så finkornede som muligt. Det vil i praksis sige, at data i det mindste skal kunne trækkes på individuelt brugerniveau. Selvom brugerfladen i Adobe Analytics Workspace er alsidig, er Adobe-data altid aggregeret. Det er ikke enkelt-brugerobservationer du får ud, men opsummering af adfærd for mange forskellige brugere som f.eks. sidevisninger for en given side, bounce- og konverteringsrater. Som udgangspunkt giver Adobe dig derfor ikke de data, du har brug for, for at kunne lave analyser fra den enkelte brugers perspektiv.

Data Warehouse er din redningsmand

Heldigvis giver Adobe dig mulighed for at trække de enorme mængder data, der er nødvendige, ud. Ja faktisk er der hele to forskellige måder at gøre det på:

1. Den første tilgang er Data Warehouse. Her kan du vælge mellem alle breakdowns (dimensioner), metrics og segmenter for et hvilket som helst foruddefineret dataområde. Disse data er allerede forbehandlet og samlet af Adobe.
2. Den anden metode er at bruge datafeeds. Her får du delvist processerede data, som er blevet sendt til Adobe. Sammenlignet med Data Warehouse er dette meget granulære hit-level data.

Jeg valgte at bruge den første metode - Data Warehouse-metoden - til at hjælpe min kunde. I resten af dette white paper fortæller jeg, hvordan jeg gjorde.

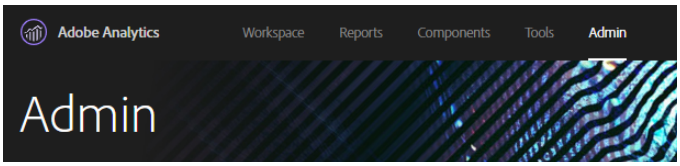
Levering med det samme. Ingen størrelses-begrænsninger: Træk Data Warehouse-data fra Adobe Analytics API med R og RSiteCatalyst

Jeg bedst kan lide at trække data ud af Adobe Data Warehouse ved eksportere dem med Adobe Analytics API og det statistiske programmeringssoftware, R. Det er hurtigt, og det giver dig meget granulære datasæt med høj volumen. Men det kræver at du kan kode. (Hvis du ikke er R-fan eller kodenørd, er der en anden – lidt mere begrænset - måde at få fat i dataene. Det kan du læse om i næste afsnit).

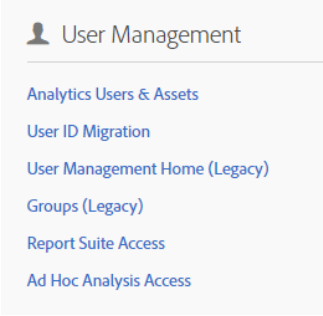
Jeg foretrækker selv at bruge RSiteCatalyst-pakken til dataudtrækket. Det giver dig mulighed for at requeste store mængder data uden brug af JSON. (Hvis du bedre kan lide JSON kan du også tilgå Adobe Analytics API'et via JSON med f.eks. Postman). R kan installeres på Linux, Windows og Mac OS X fra cran.r-project.org. Hvis du vil følge i mit fodspor med R, anbefaler jeg, at du også downloader RStudio fra rstudio.com. På den måde får du et brugervenligt og integreret udviklingsmiljø.

Efter at have installeret R og RStudio skal du sikre dig, at din brugerkonto har adgang til webservices. Bagefter skal du, for at kunne udnytte RSiteCatalyst, identificere dig med dit brugernavn og password (secret) i Adobe Analytics:

1. **Gå til Admin.** Klik på "Admin" i Adobe Analytics.

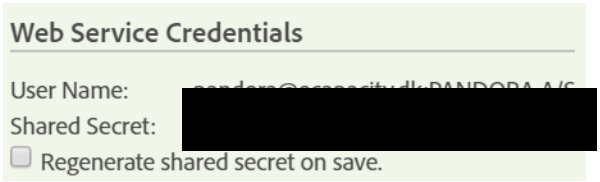


2. **Gå til "Analytics Users & Assets".** Klik videre "Analytics Users & Assets" sektionen:



3. **Find dig selv.** Brug søgefeltet til at finde dig selv:

4. **Connect til API'et.** Klik på dit "USER ID". Dette åbner et view med de detaljer, der er forbundet med din brugerkonto. Under "Web Service Credentials" overskriften står de credentials du skal bruge for at connecte til API'et. "User Name" er din email-adresse efterfulgt af virksomhedsnavnet. "Shared Secret" er en streng på 32 karaketerer.



Nu hvor du har fundet dine credentials, kan du forbinde til API'et med R ved at bruge scriptet herunder, der installerer og loader RSiteCatalyst i R:

```
# install package if required
if(!require(RSiteCatalyst)) install.packages('RSiteCatalyst')

# load package
library(RSiteCatalyst)

# fill in the user name and shared secret obtained
[redacted]

# handshake with the API
SCAuth(key, secret)
```

5. **Request report suites data frame.** Nu hvor der er skabt forbindelse til API'et kan du begynde at sende forespørgsler. Start med at requeste en data frame som indeholder den report suite du vil udtrække data fra

```
# request data frame that contains available report suites
report_suites <- GetReportSuites()
```

6. **Gem som vector.** Nu kan du åbne en data frame som indeolder report suite ID'et under "rsid" headeren. Report suiten som du vil eksportere data fra, kan nu gemes som en character vector:

```
# store the desired report suite as a character vector
report_suite <- 'myreportsuite'
```

7. **Request elements, metrics og segment data frames**. Du har nu defineret hvilke report suite der skal eksporteres data fra. Næste skridt er at requeste data frames som indeholder alle de relevante elementer (dimensions), metrics, segmenter, props og eVars:

```
# request data frames that contain elements, metrics, and segments
elements <- GetElements(report_suite)
metrics <- GetMetrics(report_suite)
segments <- GetSegments(report_suite)
props <- GetProps(report_suite)
evars <- GetEvars(report_suite)
```

8. **Tilknyt ID'er.** Hverken Analytics Visitor Id eller the Experience Cloud ID er indholdt i elements data framen. Derfor skal vi tilknytte disse dimentsioner til elements data framen:

```
# append analytics visitor id to elements data frame
elements[nrow(elements) + 1,] = list('visitorID', 'Visitor Id', NA, NA, NA, report_suite)

# append experience cloud id to elements data frame
elements[nrow(elements) + 1,] = list('marketingCloudVisitorID', 'Experience Cloud Visitor ID',
NA, NA, NA, report_suite)
```

9. **Specificer headers.** Nu kan vi tilgå data frames og specificere hvilke items vi vil bruge i rapporten. Det kræver, at der refereres til alle items med deres værdi i "id" headersne. Herunder specificerer vi de headere der skal bruges for at kunne reequeste antallet af page views og average time spent on page, brudt ned på visitor ID, page name og device:

```
# specify headers that are to be used in request
used_metrics = c('pageviews', 'averagetimespentonpage')
used_elements = c('visitorID', 'evar1', 'mobiledevicetype')
```

10. **Få navnene til at korrespondere.** Som nævnt, har"id" headers ikke altid meningsfulde navne. "Evar1", f.eks., repræsenteter page name. Heldigvis har "id" headeren også en korresponderende "name" header. En reference data frame der indeholder de korresponderende navne, kan skabes med denne snippet:

```
#  create a column that includes the metric id
metrics_map <- data.frame(ids = c(used_metrics))
# create a second column that includes the metric name
for (i in 1:nrow(metrics_map)) {
  metrics_map$name[i] = metrics$name[metrics$id == metrics_map$ids[i]]
}

# create a column that includes the element id
elements_map <- data.frame(ids = c(used_elements))
# create a second column that includes the element name
for (i in 1:nrow(elements_map)) {
  elements_map$name[i] = elements$name[elements$id == elements_map$ids[i]]
}
```

Ovenstående er også nyttigt, når eksporten skal have meningsfulde headere.

11. **Eksporter data.** Nu er vi klar til at eksportere data med "QueueDataWarehouse" forespørgslen. Herunder inputter vi ni argumenter I funktionen:

a) *reportsuite.id* – report suite id stored in the character vector.
b) *date.from* – start date for the report (YYYY-MM-DD).
c) *date.to* – end date for the report (YYYY-MM-DD).
d) *metrics* – metrics specified in the "used_metrics" object.
e) *elements* – elements specified in the "used_elements" object.
f) *date.granularity* – time granularity of the report (year/month/week/day/hour), default to "day".
g) *interval.seconds* – how long to wait between attempts.
h) *max.attempts* – number of API attempts before stopping.
i) *enqueueOnly* – only enqueue the report, don't get the data. Returns report id, which you can later use to get the data.

Som default fortsætter funktionen med at løbe i ti minutter før den stopper (120 attempts adskilt af 5 sekunders pauser). Min erfaringer er, at disse defaults skal justeres opad for at kunne klare request for større eksporter.

Det er også muligt simpelthen at sætte rapporten i kø uden faktisk at modtage data ved at sætte "enqueueOnly" til "true".

Når denne snippet køres, vil der blive requested en rapport med prædefinerede metrics og elementer og det opjusterede antal forsøg og pauser:

```
# run data warehouse request
export <- QueueDataWarehouse(reportsuite.id = report_suite,
                             date.from = '2019-05-01',
                             date.to   = '2019-05-31',
                             metrics = used_metrics,
                             elements = used_elements,
                             date.granularity = 'day',
                             interval.seconds = 10,
                             max.attempts = 240,
                             enqueueOnly = F)
```

12. **Gør headerne meningsfulde.** Nu kan du mappe meningsfulde header navne til eksport data framen. Bemærk, at "datetime" altid er i første kolonne:

```
# apply meaningful header names for request
header <- c("datetime", elements_map$name, metrics_map$name)
names(export) <- header
```

13. **Excel.** Hvis du gerne vil arbejde videre med data i excel, giver R dig en let made at eksportere data frames som .csv filer:

```
# save export as excel file
write.csv(export, 'C:/Users/myname/documents/exports/export.csv', row.names = F)
```

Alt dette er, selvfølgelig, kun et eksempel på hvilken slags data du potentielt kan eksportere. I virkelighedens verden kan du eksportere data med en masse andre metrics og elementer, og transformere data så de passer til dit eget behov.

Ventetid og størrelsesbegrænsning: få resultaterne af dine Data Warehouse request pr email

Hvis du ikke har lyst til at kaste dig over R-programmering, er der en anden måde at få fat i dine granulære data. DIrekte i Adobe Analytics-interfacet kan du specificere dine rapportdetaljer og skrive din email og få data tilsendt. Hvis størrelsen på den mængde data du skal trække ud kan holdes under 10 MB og hvis du ikke har noget imod at vente lidt på at mailen med data kommer frem (det kan tage flere timer), er dette en fin metode.

1. Log into Adobe Analytics…
2. Hover over the "Tools" header and click on "Data Warehouse".
3. Specify the "Request Name". This is done in order for you to locate your request in the "Request Manager" afterwards.
4. Select the desired Report Suite you want data from in the top right corner.
5. Select either a custom or preset "Reporting Date".
6. Select the level of granularity.
7. Select one or multiple segments from the "Available Segments" list (not mandatory). To select multiple segments, hold ctrl and select the desired segments.
8. In the "Breakdowns" section three categories are available:
    a. *Standard:* This contains all out-of-the-box dimensions that you can find in Workspace. Importantly, you can choose the Visitor Id and the Experience Cloud Visitor ID, often a prerequisite when applying ML algorithms.
    b. *Custom:* This contains all eVars and props that are available in Workspace.
    c. *Segments:* The data warehouse request can also be broken out by segments
9. Similarly, in the "Metrics" section two categories are available:
    a. *Standard:* Again, this contains all out-of-the box metrics that you can find in Workspace.
    b. *Custom:* This contains all events and instances of eVars.
10. Type the email address you want the report delivered to and the "File Name".
11. Schedule the report to be sent immediately, monthly or yearly.
12. Click "Request this Report" to start scheduling the report.

Nu skal du bare vente på at rapporten ankommer til din indbakke (hvilket, som noteret ovenfor, kan tage lidt tid).

Held og lykke med at finde "dine" sandheder

Jeg gik ad "R-vejen" for at hjælpe vores kunde med at undersøge hvor meget skade de lange loadtider på hans site egentligt gjorde. Jeg kunne også have brugte den anden metode – den med email. Men i mit tilfælde ville størrelsesbegrænsningen og ventetiden have været træls.

Hvis du selv vil lave noget tilsvarende af det jeg har gjort, er begge metoder fine – og hvis du gennemgået alle skridtene ovenfor, ender du med et fint, finkornet dataudtræk fra din Adobe Analytics. Et udtræk, som bare venter på, at du selv går i gang med din egen avancerede analyse, finder spændende mønstre i data og forudsiger hvilke handlinger fremtidige brugere af dit website med størst sandsynlighed vil udføre. Held og lykke med at finde "dine egne" sandheder.

Om eCapacity: We empower you to grow your digital business

eCapacity er en ledende, dansk baseret digital rådgivervirksomhed med tre fokusområder: Vi hjælper med at udarbejde og implementere ***digitale vækst-strategier***. Vi hjælper med at finde ***indsigter i data***, og udnytte dem, så de kan forbedre din forretning. Og vi hjælper dig til at få det maksimale ***udbytte af dine digitale platforme***.

Sådan giver vi værdi

Vores kunder sætter typisk pris på os, fordi de oplever, at vi giver dem værdi med:

- Digitale forretnings- og data-specialister, der hjælper med at udvide den digitale slagkraft.
- Et stærkt kommercielt fokus, og løsninger, hvor de vigtige forretningsmæssige målsætninger bæres helt igennem
- Bred branche-erfaring med digitale strategier, databaserede indsigter og platforms-udnyttelse fra en lang række store danske og europæiske virksomheder

Vi arbejder for kunder som Pandora og Sky, Velux og TV2, Novo Nordisk og Nykredit samt mange, mange flere.
Og vi arbejder typisk sammen i lang tid.

Hvad er din udfordring?

Vi vil meget gerne høre om dine udfordringer og give vores bud på hvordan vi kan hjælpe.

Kontakt Per Rasmussen (pr@ecapacirty.dk, +45 20 42 29 52)
Nicolai Porsbo (np@ecapacirty.dk, +45 60 78 19 00)
Andreas Petersson (ap@ecapacity.dk, +45 51 71 43 63)
Martin Wammen (msw@ecapacity.dk, +45 25 75 75 9) direkte.

The eCapacity team. Per Rasmussen, Sara Pedersen, Kristina Vehn, Nicolai Porsbo, Katrine Tølbøll Jacob Bay, Troels Moltsen, Mellody Sol, Andreas Petersson, Jacqueline Feuerborn, Donal Taylor, Elli Simonen-Andersen, Mirella Panduro, Jon Lund, Anja Kjær-Christensen, Ina Rosen, Laura Perge, Jeppe Dahl-Østergaard, Karen Rasmussen, Jacob Boye, Michael Strader, Nina Andersen, Sem Blaksmark, Frederikke Hansen, Rie Bagge-Nielsen, Henrik Mortensen, Kasper Andersen, Martin Wammen, Christopher Krolmark, Martin Munk