Advanced Attribution modeling on CLD with Essentia

A client of Mediamind wanted:

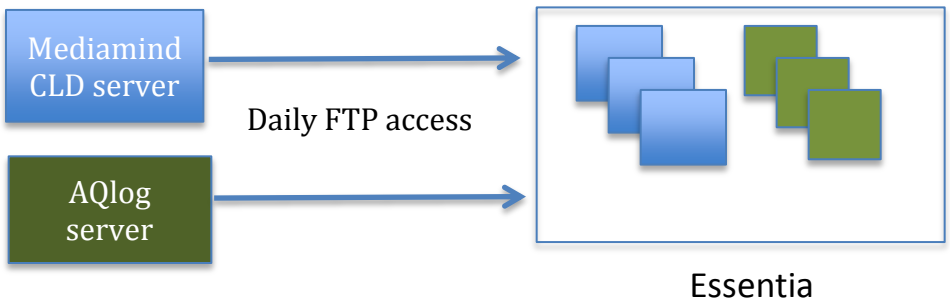
1. Campaign analysis report associated with search keywords
2. Advanced Attribution modeling, very custom logic to meet specific requirement.

We provided to the client

1. AQ-tag. A tracking tag, (similar to Google Analytics) to get all web logs.
2. Essentia: Data Management and Processing
   - to manage CLD (Cookie Level Data) downloaded from MM server
   - to blend CLD and AQlog by correlating cookies
   - to create Customer Journey
     (concatenate, join, multiple logs into single journal data)
3. Apply many different Attribution Models

Managing CLD and AQlog on daily basis

Essentia collects log files via FTP/SFTP daily from Mediamind CLD server and AQlog server and store them into each folder for each client.



Blending CLD and AQlog to create CJ (Customer Journey)

ETL (Extract Transform Load)

CLD are in zipped format. They could also be nested zip. CLD also has headers.

First step to analyze CLD is unzip data and clean up. So, ETL is the typical first step to deal with CLD. However, it's not easy to deal with large number and large volume of CLD with any existing ETL tool. Essentia does not need additional ETL tools. It can process CLD directly. No extra time to uncompress zipped files. No extra workspace to hold huge unzipped log files.

Intelligent Cookie Correlation (fuzzy matching)

Cookie of CLD and AQlog are completely different but they may belong to the same user. Correlating cookies between different logs could be very difficult because keys are typically fuzzy.

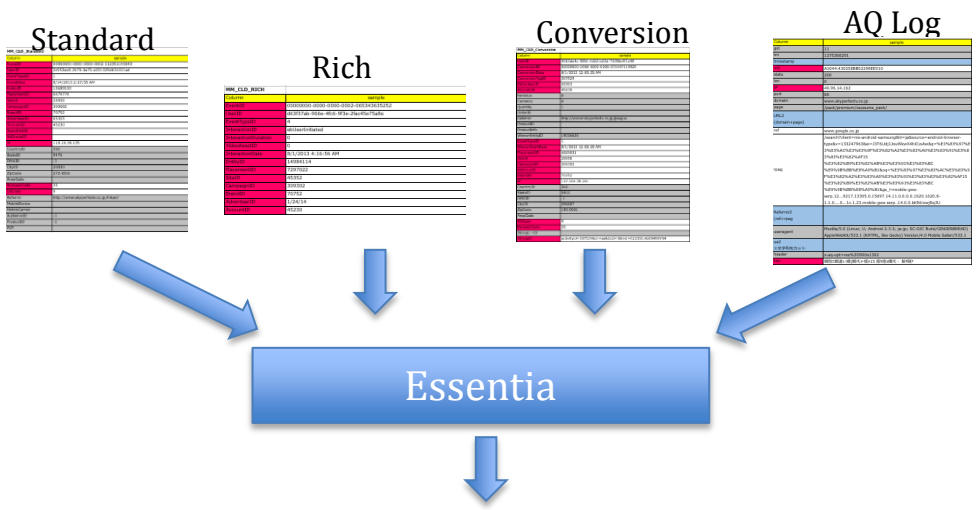
For instance, IP address and Useragent and timestamp could be combined key to identify the same user between two logs. However, timestamp could be a few seconds off randomly. JOIN or Lookup does not work to correlate cookies, which is where you need to use intelligent fuzzy matching.

Essentia can support very complex and intelligent matching logic to correlate cookies. It works very fast and makes correlating cookies buried in billions of log events possible.

Blending multiple log into Customer Journey

Filtering, transforming, concatenating, Joining and sorting process applied to create Customer Journey Table.

Customer Journey Table is a journal of all events associated with a user. When and which Ad finds user at where, when and which Ad user clicked at where and also when and which search engine and keyword user used. Also when user gets converted.



Customer Journey Table

"date"
"UserID":
"BrandID"
"BrowserCode"
"CampaignID"
"ConversionTagID"
"EntityID"
"EventTypeID"
"InteractionDuration"
"InteractionID"
"OSCode"
"OrderID"
"PlacementID"
"ProductInfo"
"SiteID"
"Source" ---- Source (1- Standard 2: Rich 3: AQ log 4: Conversion)
"url"
"rurl" --- referrer
"key" --- search keyword
"date_diff_sec" -- time difference from previous event
"all_CV_FLG" --- conversion status flag
"score1" --- Equal distribution score
"score2" --- imp:click = 1:20 weighted score

Attribution scoring

Multiple and custom attribution modeling strategies

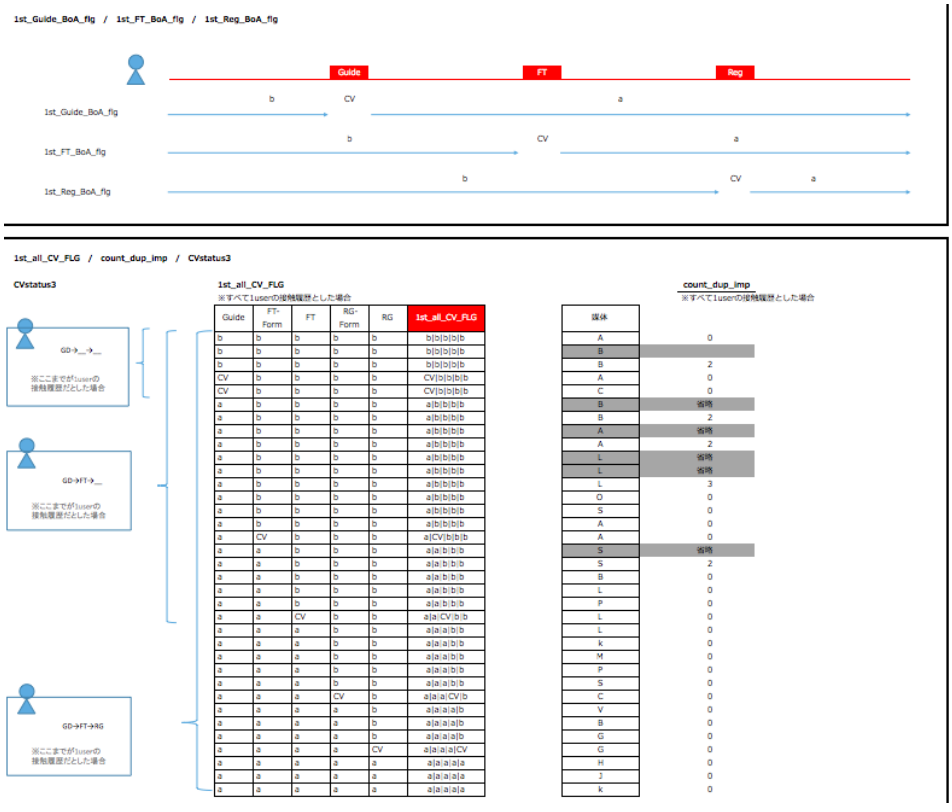
There are multiple conversion points. The client wanted attribution modeling for each different stage.

For example

> Conversion1 = landed on campaign page, conversion2 = Subscribed for free trial and conversion3 = purchase.

Client wanted know which Ad contributed most for cluster of visitors who have already viewed the campaign page but have not yet subscribed to the free trial. Etc.

This is typically a very tough requirement to do such intelligent and custom processing over billions of logs. However, with Essentia we could fullfill the special requirement quickly.



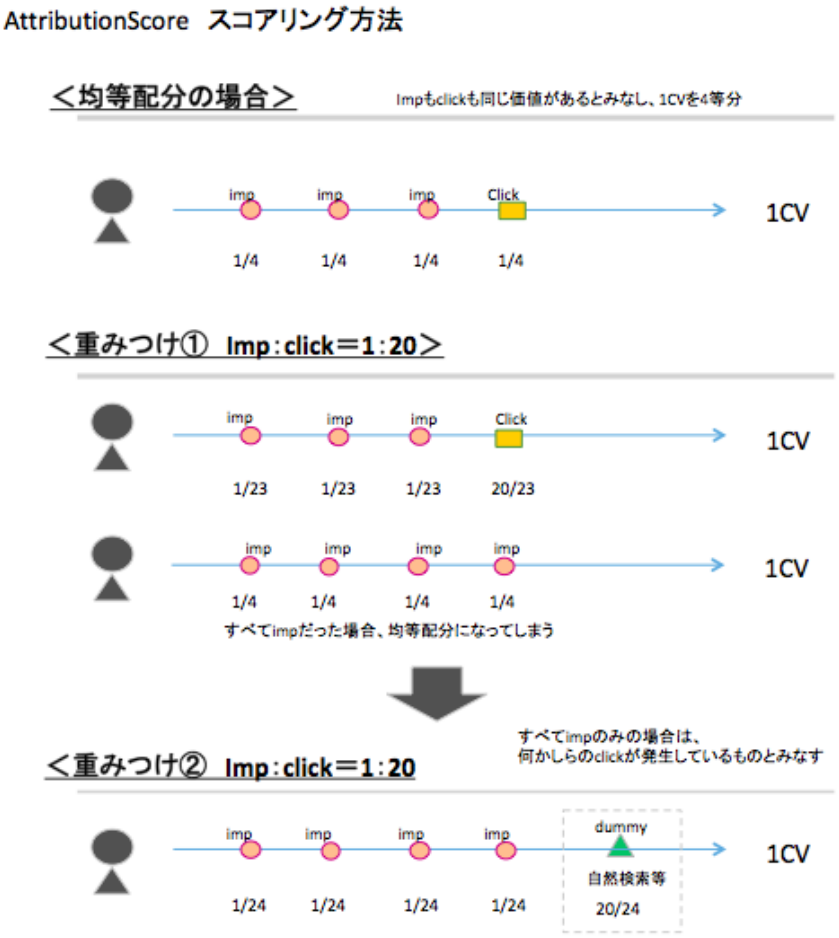
Multiple scoring algorithm

There are several different algorithms to score, such as Last touch/click, First touch/click, equal distribution, time decay and weighted scoring. Client wanted to have all of them to explore the best.

One example of the customer requirement shown below.

0. Equal score distribution.
1. Weight click 20 times more than impression.
2. If user did not have click event consider as if there were a virtual click then weight click 20 times more.

Essentia allows us to apply any advanced algorithm over billions of logs very easily.



Sample result

Calculate attribution score for user between CV1 (landed at campaign page) and CV2 (Free trial), and roll up per combined key of SiteID, PlacementID and EntityID. 2 different scores were calculated from different algorithms to compare.

siteid	placementid	entityid	score1	score2
45352	7297026	15000492	0.17	0.10
45352	7297025	15000490	0.28	0.15
45352	7297024	15000466	2.45	0.81
45352	7297023	15000461	0.31	0.07
45352	7297022	14984114	2.18	0.65
33353	7080864	14510939	0.50	0.05
33353	7080860	14510918	0.50	0.05
19534	7080571	14510937	1.00	1.00
33354	7022362	14445509	0.32	0.12
33354	7022361	14445428	0.46	0.12
33354	7022359	14445315	0.03	0.02
33354	7022494	14442982	0.92	0.88
33354	7022493	14442407	0.22	0.09
33354	7022492	14442384	1.11	0.93
45492	6994835	14359072	3.72	4.19
45492	6994832	14359068	1.00	1.00
45493	6994744	14358789	1.00	1.00
45493	6994652	14358022	2.83	3.29

Performance

Essentia is a very scalable cloud based solution. It is currently available at both AuriQ colo and Amazon AWS (can be deployed anywhere).
We have several case studies to show the performance.

Client A: Advanced attribution scoring over monthly data (about 1 billion logs). It used to take days to process to calculate score with a conventional solution. Essentia on AWS (2 node) complete the whole process (from zipped CLD to the attribution score table) in less than 1 hour.

We have many other case studies including a case to analyze 30 billion logs archived in more than 40 thousand gzip files within an hour utilizing 50 EC2 nodes.