SEVEN CRUCIAL BUILDING BLOCKS

BUILDING A REAL-TIME CUSTOMER DATA PLATFORM

PREPARED BY
ONMARC

ONMARC

SEVEN CRUCIAL BUILDING BLOCKS

BUILDING A REAL-TIME CUSTOMER DATA PLATFORM

When a random organization is asked if their Customer Data environment is ready to use in an ideal omnichannel era, what do you think the honest answer will be?

The term real-time CDP (Customer Data Platform) is relatively new. So far organizations have expanded on legacy CRM systems and databases to provide similar functionality. In our view there is need for a new approach. In our vision seven building blocks need to be in place and in line to create the ideal CDP.

A CDP's primary function is to serve as a unified data source for all data about an organization's identified audience including (ex-)customers, prospects and incidental visitors. It typically serves two main application areas:

- **Insight: using data to learn about all individual actions and interactions both for analytical and intelligence purposes**
- **Engagement: using individual, extended profiles to better serve the audience on an individual level, preferably one-to-one**

- by Vincent Kooijman

PREPARED BY
ONMARC

ONMARC

SEVEN CRUCIAL BUILDING BLOCKS

BUILDING A REAL-TIME CUSTOMER DATA PLATFORM

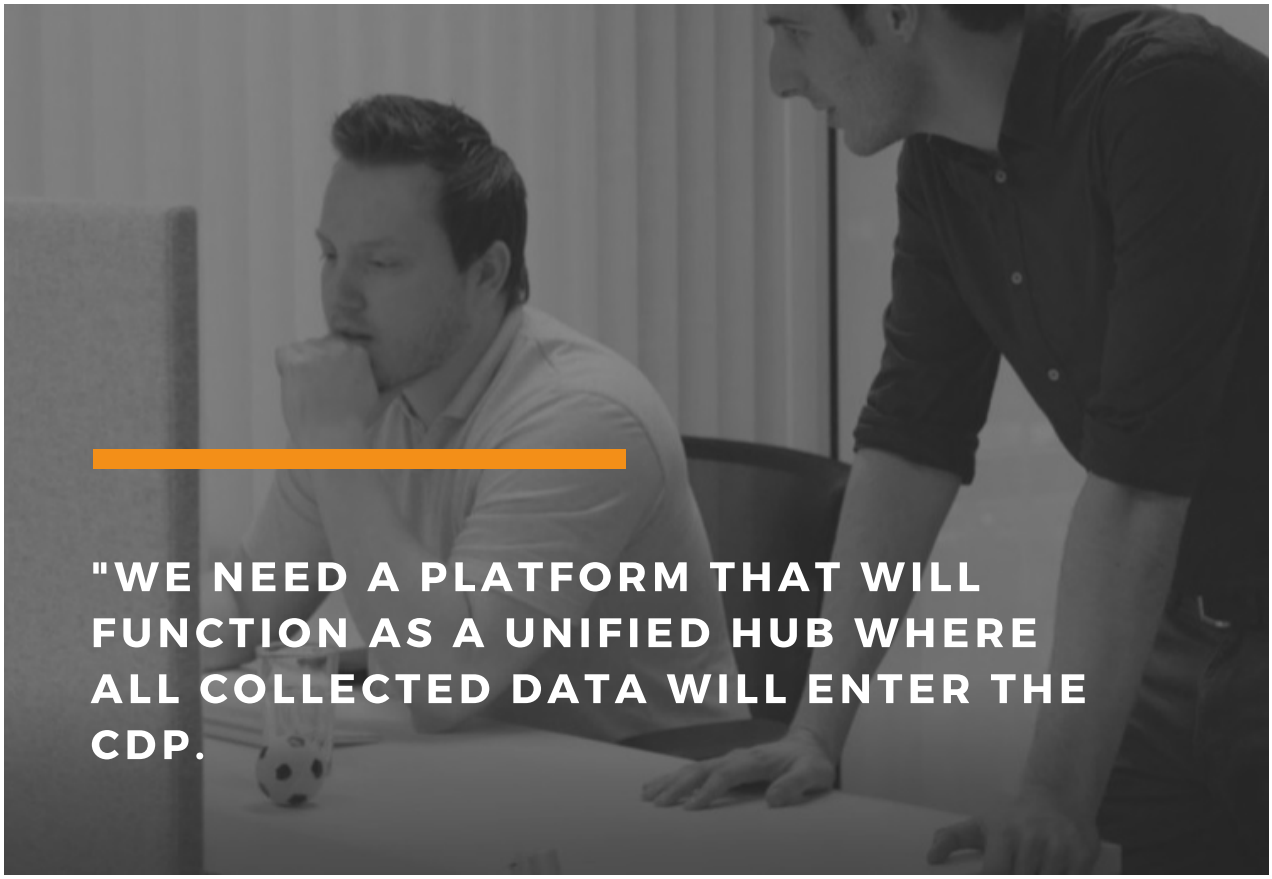
- by Vincent Kooijman

VINCENT KOOIJMAN
MANAGER DEVELOPMENT

Vincent Kooijman is managing OnMarc's development team. OnMarc offers services to assist customers in the capturing and use of (online) customer interaction. Vincent is an inspired developer and expert in the area of Java and Hadoop applications in MarTech environments. He is responsible for the development of the OnMarc Customer Data Patform. This platform enables true omni-channel marketing by integrating a versatile set of digital data sources.

Why a Customer Data platform in real-time?
A modern CDP has to operate in real-time. Driving personal communication using modern Content Management Systems or marketing techniques such as Programmatic Advertising requires behavioral and transactional data to be available instantly as otherwise communication may become irrelevant or money is spent without any purpose.
The real-time nature of a CDP creates new challenges for both developers and operations engineers. In a batch oriented environment, it was easy to have intermediate datasets and to re-process data, in real-time we must take a more resilient and agile approach.

SEVEN CRUCIAL BUILDING BLOCKS



1.

DATA COLLECTION

A CDP collects data from many different sources. Data from digital channels such as websites and apps, social media and internet-of-things, paid marketing media and customer feedback systems all need to be integrated with back-office data from CRM systems, customer service and call centers. Other supplemental data sources with valuable information about the circumstances under which the audience makes decisions such as weather data, economic climate data, crime rate data, house- and car-for-sale data may also be integrated.

But not all data is available in real-time. Some sources generate data in real-time, some of them only on an hourly or daily basis. This means we need a platform that will function as a unified hub where all collected data will enter the CDP, no matter whether the source system generates this data in real-time or batch. This will provide us with an effortless way of processing the data, without having to worry about when and how the source system generated the data.

SEVEN CRUCIAL BUILDING BLOCKS

At OnMarc we have chosen to use Apache Kafka for this. Kafka is a streaming platform which can handle millions of events per second. Kafka is the perfect choice for us due to the large volume of data we handle. Kafka receives or collects data from remote systems and stores the data in message logs. Data in Kafka is structured into 'topics'. At OnMarc, each topic represents an individual type of event. For instance, all Tweets will be directed into their own topic, and mouse-overs on your website into another. This makes it easier for downstream applications to select the data they require.

For each data source we integrate into the CDP, we have created a producer that processes data from that source as soon as it becomes available, real-time or not. The producers will structure the data into events and send them to the various topics in Kafka.
The benefit of this approach is that we only have to manage one single flow of data from this point onward. Each producer has the responsibility of knowing when data is available and then processing it depending on its source.

One element that is crucial here: this approach makes Kafka the heart of our data processing environment.

SEVEN CRUCIAL BUILDING BLOCKS

2.

INTERNAL USAGE AND DISTRIBUTION

So, now we have all data available in Kafka, how can we make this data useful for further processing? We have to consider that:

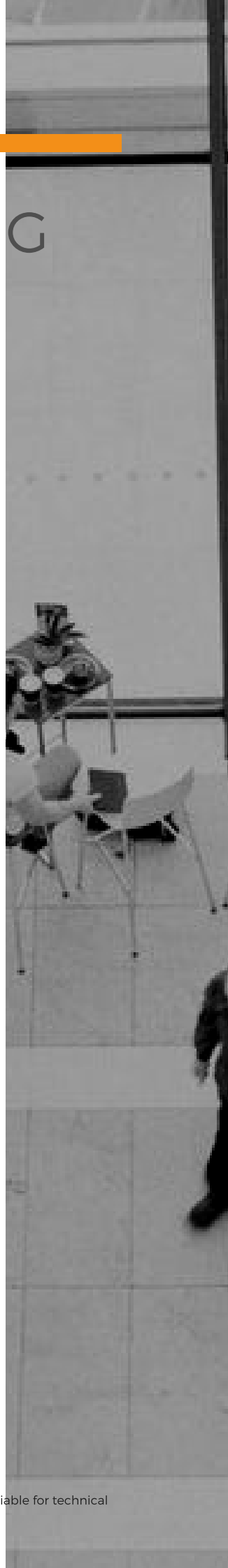
- Data for each topic is only structured in its own context, and may not be ready for the various applications expecting a more uniform data structure.
- Data remains in Kafka for a limited time. Remember Kafka serves primarily as a queueing mechanism.

For further processing of data Kafka uses the concept of consumers. A consumer is an application that subscribes to specific topics within Kafka. Whenever events for those specific topics arrive they will be available for the consumer to process them in real-time. These consumers will process the incoming data for a specific use-case within the CDP. For instance, we might have a consumer to predict churn behavior by analyzing mouse-over and click data.

Kafka consumers may also need to combine different events, maybe even from different sources. Here special attention to synchronization is required. As Kafka topics receive and store data independently and asynchronously, it is necessary for consumers to take care of this. An example: when a visit to your website starts, an event identifying the session and user is emitted. When the visit ends, an event signaling the end will be created. What happens if, for some reason, an end event arrives before the start event? Consumers relying on this data need to take that into account when processing data from different sources, or their processing logic might fail. The consumer will have to make the right decisions and combine the input streams of all events into more logical structures that can be used within downstream CDP applications.

Defining a more logical structure is an important role for the Kafka consumers. Data within Kafka topics are independent of each other and have no real pre-defined structure. The data structure within these topics will most likely differ when originating from different sources.

We have created consumers as micro services that take care of processing the data for a specific source. . An example is a service that will gather data for social media related events and combine them into a single social media record.



SEVEN CRUCIAL BUILDING BLOCKS

This record will be structured into a format our other applications expect and understand. The record will then be put back on Kafka, to be processed further.

Finally, OnMarc has combined several of these services into three applications that serve the following purposes:

- Analytics and intelligence using an Impala SQL-environment on top of a Kudu file system
- Structuring data into a Data Management Platform for Programmatic Advertising
- Delivering data to 3rd-party applications using a real-time push API

3.

RESILENCE AND STABILITY

When operating in real-time, special attention needs to be given to the resilience and stability of the platform. However, systems fail and we need to design our system to tolerate failure as:

- Kafka is the heart of the system, it requires non-stop availability. Kafka downtime must be avoided as most producers have a limited internal buffer capacity and will start dropping data when Kafka is not available.
- Due to the limited buffering space available in individual components, we need to make sure that data is processed at a sufficient rate to handle all incoming data. Losing data because we are not able to keep up is not an option.
- Given the vast volume of data, reprocessing of data needs to be prevented, preferably eliminated. Moving the processing back in time to reprocess data is not feasible anymore.
- Internal services that our applications depend on might fail. We need to guarantee data is not lost due to the unavailability of individual components.

So, what can we do to protect the environment against data loss? The most obvious one is to make most components highly available by having redundant nodes. For instance, Kafka allows for a combination of nodes in an active/passive scenario. Protecting this crucial component against hardware or system failure is the first step, even though Kafka itself has proven to be a very reliable technology.

The second important item is allowing for sufficient buffer space. Due to the fluctuations in traffic volume of most digital channels (campaigns, seasonal influences) it is important to have sufficient buffering capacity.

SEVEN CRUCIAL BUILDING BLOCKS

"WHEN OPERATING IN REAL-TIME, SPECIAL ATTENTION NEEDS TO BE GIVEN TO THE RESILIENCE AND STABILITY OF THE PLATFORM."

When using Kafka, one can opt for a certain lifespan of data to be retained within the topic or a fixed size in storage capacity. The first strategy guarantees consumers with ample time to process data, but potentially requires a lot more storage capacity due to unexpected increases in traffic which might emit an unexpected amount of data within a certain time period. The second option guarantees a maximum and predictable amount of storage usage, but does not guarantee all consumers can process the data in time in case of a spike in traffic. Data might be dropped due to storage limits before all consumers had a chance to handle the incoming data.

When operating in batch mode (as opposed to real-time mode) it is easier to schedule reprocessing of data. In real-time mode either the real-time process must be stopped and historical data has to go through the same process again starting at a much earlier point. This will cause the customer profile data to be out of date, and therefore personalization will be affected . An alternative option is to process the 'older' data once again in parallel with the real-time process. This process is more complex and requires substantially more resources in order to protect the real-time processing.
In any case, preventing the system from having to reprocess data is a much better alternative. However, this requires better control over the oriFinally, as the volumes in CDPs typically are huge we have to think about protecting stored, processed data in databases.

SEVEN CRUCIAL BUILDING BLOCKS

Finally, as the volumes in CDPs typically are huge we have to think about protecting stored, processed data in databases. As described above, reprocessing source data is a difficult and a costly mechanism. Reprocessing many months or even multiple years of source data can be a very long, if not indefinite and impossible process. Backing up many terabytes of processed data is also difficult, for two reasons: sheer volume and privacy and compliance regulations. With the GDPR regulation it must be possible to remove data upon request by a user. But how will that be done when multiple backups contain data for that user?

When using a distributed storage environment like HDFS or Kudu, data will be replicated over multiple nodes. This gives us a sufficient amount of redundancy for the stored data and will protect us against component or system failure. But beware, this does not protect us from human error.

Building a reliable and resilient system that guarantees a high rate of availability and little to no data loss is not an easy task. It involves careful fine-tuning of individual components for your specific use-case.

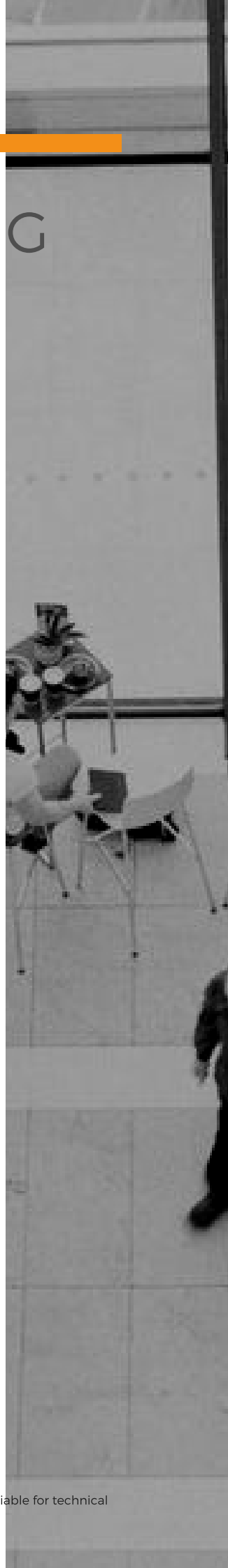
4.

DATA RETENTION

A CDP contains a vast amount of data from many sources. An incredible potential for analysts and a great source for individual personalization. But how far do we go back? Do we keep data forever or do we keep a compact data set available? Storage is cheap, so the answer probably lies within the business. But there is a legal constraint to consider as well.

In the past life was simple. For legal reasons companies had to secure their data for a number of years (typically 5 or 7 years). This was for auditing and control reasons. With the new privacy and compliance rules as set forth in the GDPR, data related to an individual can only be kept as long as it is relevant for the business use-case the data was originally recorded for. This typically is far less than 5 years.

And then there is the analytics balance. Where on one hand It is useful to have as much data as possible, the numbers can get quite big and slow down analytics.



SEVEN CRUCIAL BUILDING BLOCKS

BUILDING A REAL-TIME CUSTOMER DATA PLATFORM

Moreover, data going back several years might no longer be relevant. Your website, business and customers will have changed over time. When we record billions of data points, one needs to balance between wanting to retain all data, maybe even the irrelevant, and sheer processing power available to give a more accurate analysis using the larger dataset.

5.

IDENTIFY CONTROL AND MATCHING

When integrating data from so many different sources, special attention needs to be given to identity management, one of the most critical components of a CDP. The digital channels in particular operate with a wide variety of identifiers: cookies, sessions, clicks, impressions, customer numbers, e-mail addresses, devices and more. Add to that identifiers from call centers and the back-office, data that is structured, but typically not directly matched to ids from digital channels and a challenge has just arisen.

ONMARC

- OnMarc helps to identify customers individually and as a group. We capture all customer interaction across all digital touchpoints, which will help you recognize patterns and predict choices, annoyances and client requirements. With the use of OnMarc solutions, customer experiences are personalized, conversions increased and brand experience satisfaction maximized.

SEVEN CRUCIAL BUILDING BLOCKS

When data is received in real-time it needs to be linked in real-time as well. The most likely data that is available in real-time is the data from the web- and mobile app-environment. Using this identifier structure as a foundation has additional benefits as this environment provides identifiers not only to customers, but also to unknown visitors.

Where most identification can be done directly as two sources use a common identifier, in certain instances it may require further processing. For instance, when content must be scanned for information such as e-mail addresses or zip code/house number info. Some data offers a high certainty about the individual (i.e. customer number) when matched, where others may identify a household or a small group of people (i.e. e-mail address, zip-code).

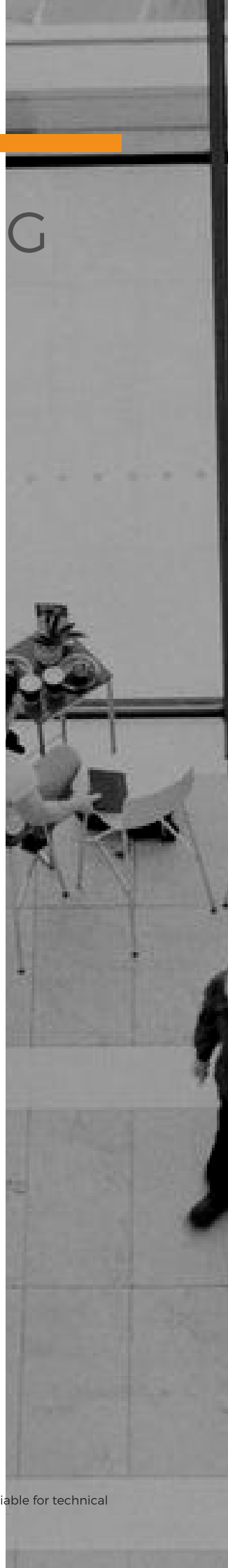
In any case it is useful to use upfront identification as personalization depends on it. So, real-time and high-quality identifier matching is key in the use of a CDP.

6.

THIRD-PARTY INTEGRATIONS

Having a CDP is not an objective on its own. It is a means to get better insight into a company's audience and it helps personalize individual communication. However, the true value lies in how many different systems it can connect with allowing data to come in and data to go out. Incoming data is mainly described above in the section about data collection. Also, with machine learning and real-time analytics solutions, an extra element is required. Users of the CDP need to be able to integrate the CDP with other solutions, such as modelling software to provide various scoring values against visitor profiles based on individual behavior.

So, data going out is equally important as data collection as this will generate the value a CDP brings. For analytics and intelligence functions data can be accessed by using default connection mechanisms allowing most applications to access the data when stored within a relational database like Apache Impala.



SEVEN CRUCIAL BUILDING BLOCKS

For customers that deploy a real-time decisioning or next-best-action system for omni-channel personalization the CDP is the most important source of information. The CDP needs to trigger these systems by pushing data as and when needed but also provide a full, up-to-date and real-time profile to these systems. For programmatic advertising and less complex online personalization, anonymous segments must be provided to external software environments. This can be DSPs for programmatic advertising, recommendation engines, but also content management systems.

Any external application that wants to utilize data from the CDP will require access to the dataset. These applications need to be able to request specific bits of data as needed by utilizing various APIs, for instance a REST API.

Again, the value of a CDP lies not only in the data it collects, but just as much in its ability to have other systems access and reuse the wealth of data available.

7.

PRIVACY, SECURITY AND COMPLIANCE

Having a CDP as a single source of data has many benefits, also for privacy and compliance, something you might not realize at first. The CDP is under direct management of the organization, which means it will be in full control of the nature of the data being stored, and for how long it will be retained. Moreover, data can be shared in a more controlled way with 3rd parties and typically in an anonymous fashion. This protects confidentiality of the communication between the organization and its audience.

But it also comes with great responsibility towards the customers and visitors. As a CDP may contain confidential information it must be protected against unauthorized use. Not only because of GDPR and other legal or compliance restrictions as explained above. It is the trust between a company and its external audience that is at stake!

SEVEN CRUCIAL BUILDING BLOCKS

BUILDING A REAL-TIME CUSTOMER DATA PLATFORM

An extensive authentication and authorization implementation is required ensuring (a) that only staff members with the appropriate privileges are entitled to access specific sets of data and (b) that they can only execute operations that match the user's role. Integration with a corporate access system and protocol such as LDAP is a mandatory feature. Audit trails detailing who had access to what data are also a necessity.

A secure, single data source within the CDP is a big improvement over segregated data sets that operate independently within different external organizations and with multiple and often different security mechanisms, if any at all.

CONTACT ONMARC HERE: +31 30 636 39 70 OR INFO@ONMARC.NL