Estudio de datos recolectados por Analytics en un negocio online web.

Jordi Castillo Mas

Resumen– El objetivo de este artículo es exponer el estudio realizado a Google Analytics en una empresa online de bodas. Donde se cotejará en detalle, los datos de las trazas de usuario recolectadas en un año para compararlas con un sistema local de recolección de trazas. Donde finalmente se realizará una importación de los datos de Analytics a la colección de datos local. En este estudio, se hace un análisis de los datos para ver que registros son los que se guardan en cada una de las dos colecciones de datos, y se comparan entre si para ver qué diferencias hay en cada una. El estudio se ha realizado en una compañía de servicio de bodas online, que ofrece al cliente la posibilidad de contactar con diferentes empresas del sector. Es por esta razón, qué se analizará la interacción de los usuarios en estas empresas, situando el foco del estudio en saber las visitas que realizan estos usuarios en ellas. Al final de este artículo se muestra, el porqué de los distintos datos en cada una de las colecciones; la solución que se ha dado en el estudio para poder cotejar con confianza los datos; demostrar que los datos en ambas colecciones son distintos, concluir con una estadística de los datos de cada una de las colecciones y del uso que se le va a hacer.

palabras clave– Universal Analytics, Google Analytics, Data Analysis, Compare data, Data mining, Big data, BigQueries, MySQL, MapReduce.

Abstract– The aim of this article is to expose the study carried out to Google Analytics in an online wedding company. This study will check in detail the data of the user traces collected for one year ago, comparing this data with a local datebase which have the same trace collection. In this study, we make an analysis of the data to know which records have been saved in both data collections, and compare one to other to get the differences between collections. The study was realized at an online wedding service company, that offers to the costumers the possibility to contact service with different companies in the sector. That is why, we want to analyze the users interaction with companies, focusing the study to know the users visits on this pages. At the end of this paper we'll explain: the facts that makes the data collections be different; the answer that has been rather in the study to be able to check the information of the collections; we'll demonstrate that the data in both collections are different and, we'll conclude with a statistics of the data of each collection and the uses that we will do with them.

Keywords– Universal Analytics, Google Analytics, Data Analysis, Compare data, Data mining, Big data, BigQueries, MySQL, MapReduce.

1 Introducción

Este artículo trata de describir el trabajo realizado en un proyecto de empresa privada, donde se quiere abordar un problema de minería de datos. Este proyecto consiste en cotejar los datos que se han ido recolectando durante más de 1 año en una empresa web del sector de bodas para que parejas puedan organizar el día de su boda. Estos datos se recolectan a través de la navegación de los usuarios a medida que van interactuando con el portal web y plataforma en smartphones.

Los datos se han ido recolectando todo este tiempo con la herramienta de trazas de usuarios Google Analytics y a través de una herramienta desarrollada por la misma empresa, que graba las acciones del usuario mediante la página web. Las dos herramientas tienen el mismo objetivo, registrar toda interacción usuario-página y guardar, por cada he-

• E-mail de contacto: jcastillomas@gmail.com
• Mención realizada: Tecnologies de la Informació
• Trabajo tutorizado por: Juan Carlos Sebastián Pérez (Departament d'Enginyeria de la Informació i de les Comunicacions)
• Curs 2017/18

rramienta, la información en una base de datos distinta para tener ambas colecciones separadas.

Este proyecto surge de la necesidad de querer analizar el comportamiento del usuario en la web. Pero al realizar las primeras pruebas, se observó que las trazas de usuarios de ambas colecciones de datos, que supuestamente deberían tener la misma información, no tenían nada en común. Entonces, a través de la necesidad de saber el porqué esas colecciones no son iguales y de querer obtener una unificación en los datos para poder trabajar más adelante con ellos sin tener la necesidad de trabajar con ambas colecciones por separado.

Se quiere revisar qué información hay en cada una de las colecciones de datos y ver que diferencia hay entre ellas, pero al tener un volumen tan alto de datos hace que la tarea no sea trivial. Por esta razón, se ha decidido comparar primero solo una parte de ambas colecciones. La parte que se ha decidido escoger es en la sección de empresas de la plataforma, donde los usuarios hacen contacto con los diferentes escaparates que ofrece la página de bodas.

Finalmente, se quiere combinar ambas colecciones para poder completar un estudio con detalle de las visitas del usuario en las empresas que existen en la página.

2 OBJETIVOS

Para realizar este proyecto necesitamos cumplir una serie de requisitos para poder finalizarlo con éxito. Es por esa razón, se ha querido marcar unos objetivos, donde una vez cumplidos todos se podrá dar el trabajo como realizado.

El primer objetivo es conocer el contenido de las colecciones. Para poder lograrlo, necesitamos saber en ambas colecciones de datos que se están guardando y como gestionan las herramientas para registrar dicha información. Este objetivo nos permite tener una visión global del proyecto, entender que contenido tenemos registrado y que datos estamos recolectando de la interacción de los usuarios.

El segundo objetivo es ver que conflictos residen en cada una de las colecciones respecto la otra que hacen que no sean idénticas, ya que para poder llegar a combinar ambas colecciones primero tenemos de gestionar las diferencias que existen entre ellas. Una vez tengamos una visión de estos conflictos se podrá implementar un proceso para poder discriminar esos datos incoherentes en ambas colecciones y lograr unificar las colecciones.

El siguiente objetivo consiste en realizar un análisis con BigData en las colecciones para reducir en informes el contenido de los registros de las herramientas. De esta forma, podemos obtener de en cada una de las colecciones, las visitas que han realizado los usuarios en la página.

Otro objetivo es extraer conclusiones sobre los datos obtenidos por el análisis en BigData. Se quirie analizar los resultados obtenidos de haber hecho el estudio y contrastarlos con los que, sin haber hecho ninguna modificación, estarían almacenados en la colección. Y finalmente, ver que porcentaje de coincidencia hay entre las colecciones antes y después del análisis.

El último objetivo es ejecutar un proceso para unificar los datos de ambas colecciones, este proceso añadirá a la colección de la base de datos local la información obtenida de los resultados del análisis de BigQuery realizado en Google Analytics.

Una vez completados los anteriores objetivos, tendremos realizado el análisis de los datos en Analytics y obtendremos una combinación de ambas colecciones guardadas en la base de datos local.

3 ESTADO DEL ARTE

Este proyecto está enfocado en analizar un conjunto de datos muy grande, de un orden de 4TB de registros, y generar unos procedimientos para analizar, procesar y reducir la información en distintos informes para posteriormente hacer un estudio. Por esta razón, podemos situar el proyecto dentro del campo del BigData y del Data Minning.

Las herramientas que se van a utilizar en este proyecto son muy diversas y la combinación de cada una de ellas hace que se puedan cumplir los objetivos, por esa razón, a continuación, se va a detallar cada una de ellas.

La primera es la herramienta de **Google Analytics**[1], esta permite grabar el comportamiento y acciones que realiza una persona dentro de un dominio web. Ha sido la encargada de ir recogiendo los datos de los usuarios durante todo este periodo de tiempo. Esta herramienta contiene el módulo de expansión Google 360 Suite, que nos permite hacer un mejor seguimiento de los usuarios entre las plataformas en la página. Google Analytics tiene un módulo que nos ha permitido realizar un análisis de datos exhaustivo, este es **BigQuery**[2], y consiste en una base de datos de Google con el contenido de la colección de datos de Analytics, donde permite realizar operaciones de búsquedas de forma óptima con la capacidad de procesar un gran volumen de datos en un margen de tiempo óptimo. Es por esa razón que esta herramienta ha sido muy útil para realizar el proyecto, ya que ha sido la responsable de dar apoyo para analizar los datos y lanzar el proceso de reducción datos en un tiempo muy optimo, ya que en todo momento cualquier operación no ha tardado más de 2 minutos. Por eso solo hemos tenido la necesidad de gestionar el volumen de información que movemos por la red sin tener en cuenta los recursos de cálculo que se necesitan.

La segunda herramienta ha sido un servidor de **MySQL**[3], que tiene la capacidad de almacenar datos y realizar consultas SQL para poder gestionar la colección local de datos. En esta base de datos se realizará la migración de datos de los informes que obtengamos de Analytics. Esta herramienta no tiene tanto nivel de procesado como la de Google Analytics pero al tenerlo a nivel local es mucho más accesible y podemos tener un control total de los datos almacenados.

Para poder realizar los procesos de ejecución de análisis, reducción e importación de los datos, se ha utilizado el código de programación **PHP**. Este código se encarga de gestionar los procesos de BigData, en él se preparan los procesos que tienen de ejecutar las herramientas y, finalmente, mover los datos de una colección a otra. La necesidad de utilizar este código de programación es debido a que en el entorno de la empresa ya tienen módulos para gestionar las otras herramientas, anteriormente mencionadas. Permitiéndo la conexión a la base de datos MySQL, con la posibilidad de importar y exportar datos, y realizar consultas SQL en ella. La empresa también dispone de una API para realizar la conexión con BigQueries de Analytics con

la capacidad de realizar importaciones de datos y realizar consultas BigQuery.

4 METODOLOGIA

Para el desarrollo de este proyecto hemos utilizado un modelo basado en iteraciones. Para este se podría atribuir una metodología de tipo de modelo en espiral, aunque se haya realizado en dos ciclos del espiral, no era trivial conocer el número de ciclos que se debían realizar para resolver el proyecto, ya que por cada ciclo, se han cogido unos requerimientos con el objetivo de incrementar el emparejamiento de ambas colecciones. Se ha analizado como serán los procesos de análisis y reducción; se han creado los procesos capaces de analizar y minimizar los datos en informes; al final del ciclo se han ejecutado y analizado para ver la mejora obtenida; y finalmente repetir el mismo ciclo pero con los nuevos datos obtenidos, esto hasta lograr un índice de porcentaje de coincidencia aceptable.

El procedimiento que hemos utilizado para cada iteración ha sido el mismo, es por esa razón se ha dedicado una sección en este artículo para detallar como se ha desarrollado el proyecto. A continuación se exponen los procedimientos.

4.1. Procedimiento

El procedimiento que hemos utilizado es el núcleo de este proyecto, ya que ha sido diseñado para desarrollar el proyecto de una forme dinámica y simple. El potencial de este procedimiento es que se puede realizar en paralelo más de un estudio de patrones a la vez, para posteriormente reducirlos en un informe.

Estos pasos están hechos para que se cumplan de forma predecesora del primero al último, es decir, primero cumplir el procedimiento uno, luego el segundo y así hasta el último. Estos son los siguientes.

El primero es realizar un **estudio de los registros** adquiridos por ambas herramientas de trazavilidad, comparando entre ellas a nivel de registro los datos, y hallar registros perdidos y/o añadidos en cada una de las colecciones que no esté añadida a la otra.

El segundo es realizar un **estudio de patrones** entre los registros anotados en el primer objetivo. Este punto, trata de agrupar coincidencias entre los registros y ver si realmente siempre se cumple un patrón que hace que los datos de ambas colecciones sean distintos. Es decir, intentar ver si siempre se cumple que algunos inputs de las trazas añaden los registros en una colección y para la otra no. Este punto tiene como objetivo ver el histórico de las trazas.

El tercero es realizar una **comprobación de patrones**, donde por cada patrón que hemos hallado en el punto anterior se lo tiene que someter a prueba, y ver, si siempre se cumple que los inputs de las trazas son grabadas en una colección. Para este punto, comprobaremos a nivel práctico que las acciones de los usuarios en unos inputs son registradas solo por una de las herramientas.

Una vez realizados los objetivos del uno al tres, en el caso de que el patrón se cumpla, se realizará un **proceso de marginación** para los datos con patrones de conflictos hallados anteriormente, como este proceso es muy costoso, se pueden ajuntar más de un patrón a la vez.

Una vez hecha el proceso de marginación, el siguiente paso es **reducir** los registros a esquemas. Reducir los registros a esquemas consiste en resumir todos esos registros en conjuntos de registros para poder analizar de mejor forma y precisa los datos obtenidos.

Por último, se realizará un **estudio de los datos reducidos**, para ver cuanto han mejorado nuestra coincidencia de ambas colecciones de datos.

5 EXPERIMENT

El experimento es la puesta en marcha del proyecto, trata de realizar un primer contacto entre los datos que residen en ambas colecciones e intentar hacer una prueba de concepto de unificación de los datos. Como desconocemos completamente que similitud hay entre ambas colecciones, se decide hacer un primer estudio para ver que índice de coincidencias hay en cada una de las bases de datos.

En este ciclo se aplican las reglas de los procedimientos anteriormente expuestos, para resolver la primera iteración del proyecto, aunque en este punto del proyecto se desconocía por completo que procedimiento se debía seguir para resolverlo. Pero antes de seguir avanzando, se quiere exponer que este ciclo no llega al final, ya que como más adelante se explicará, no se pueden demostrar todas las premisas pensadas para que esta iteración se cumpla. Por eso, se forzó a los datos a que se hallaran coincidencias, lo que derivo a un error fatal en la integridad de los datos obtenidos. Por esa razón se descarta la idea de seguir avanzando con un modelo incremental en el proyecto y se pasa a un modelo por ciclos, dejando el experimento como el primer ciclo y recogiendo toda la información posible hallada en él, para poder alimentar el siguiente ciclo.

Para la resolución de este ciclo, la idea que se tenía en mente es que cada registro de una colección tiene de estar fuertemente ligada a otra de la otra colección, y si no existen coincidencias quiere decir que hay errores en los registros. Como veremos más adelante en el proyecto, esta premisa es falsa, pero en este punto desconocíamos que fuese así, ya que como más adelante veremos que no son errores los registros que existen de más en las colecciones, sino que son datos que no han sido registrados por la otra colección.

Teniendo en mente que cada una de los registros tiene que estar ligado con otro, se ejecutó un proceso que recoge los datos de un día de ambas colecciones y se compara cuanta coincidencia hay.

El resultado fue frustrante, ya que ni el 2 % de los datos cruzados coincidía, es por este hecho que se asumió que esto se debía a un error humano a la hora de realizar el proceso.

El error es debido a que el proceso de cruzar datos es muy simple. Este proceso trata de hacer coincidir las visitas hechas por usuarios en empresas en un instante de tiempo de una colección con otra. El error no podía ser muy complejo ya que en este punto no se analizaba que sección de la empresa está visitando el usuario. Es por eso, que se observó que el error residía en el instante de tiempo, ya que en el momento de registrarse los datos ambas colecciones no marcaban la misma franja horaria. También se observó que no marcaban el mismo segundo exacto, es decir que los datos pueden llegar a tener un retraso de 10 segundos en la instancia de tiempo respecto la otra. Además, se

obtuvo información acerca de este procedimiento en la que ambas colecciones no tenían el mismo número de registros, ya que para un día Analytics había recaudado cerca de los 70.000 registros mientras que los de la base de datos local no llegan a los 30.000. Esto son los resultados que pudimos aprovechar de este ciclo ya que lo que se expone a continuación son más pruebas del experimento que no llegaron a aportar nada en el proyecto debido a un mal enfoque.

Una vez realizado el análisis anterior se forzó a los datos, con una modificación en el procedimiento anterior, a que se cruzaran entre si. Para así poder lograr un mejor ratio en el índice de coincidencia de las colecciones.

Esta modificación lanza el mismo proceso teniendo en cuenta que un dato se puede relacionar con más de uno en la otra colección y tener un margen de error en la instancia de tiempo de un registro de una colección con la de la otra colección, se ejecutó una prueba de concepto para diferentes índices de error: 10 segundos, 1 minuto, 15 minutos y 1 día. Los resultados fueron sorprendentes, ya que el índice de coincidencia se disparaba mucho. Pero como ya veremos este índice de coincidencia es **falso**. A continuación, se muestra una imagen con los indices de coincidencia erróneas hallados con este procedimiento modificado, las líneas verdes son el porcentaje de emparejamiento de los registros de la colección Analytics y los azules de la colección Local.

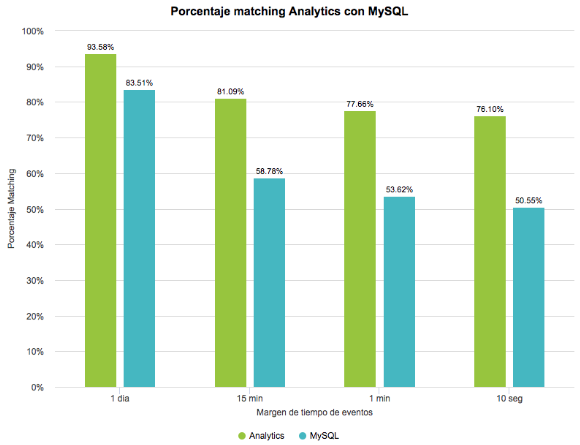

Fig. 1: Tabla de índice de coincidencia falsa del proceso de modificado del experimento.

La razón de que este proceso modificado sea falso es debido a que no se contempla que secciones está visitando el usuario; también es falso que un dato de una colección se pueda relacionar con más de uno de la otra colección y; finalmente, también es falso que un dato pueda tener un lapsus de tiempo de 1 día entero de diferencia con la otra colección. Se descartan estos resultados ya que peligra la integridad total del proyecto si seguimos con estas directrices.

Es tan grande la incertidumbre de los resultados del experimento que finalmente se da por concluido debido a que se encuentra en un callejón sin salida, ya que este procedimiento no es capaz de relacionar los datos correctamente que a la vez sean fiables. Los resultados que obtuvimos con el experimento demuestran que ambas colecciones aún tener aparentemente las mismas trazas no son iguales, por lo que intentar realizar una comparación a escala de registro resulta muy difícil.

Antes de finalizar con este ciclo se decidió que está no era la forma correcta de atacar la problemática del proyecto, por este motivo se decidió realizar una exportación de ambas colecciones de datos reducidos en un informe mensual para agrupar el número de visitas de usuario en empresas por mes. De esta forma, se desarrolló la idea de iteración por ciclos, intentando atacar el problema por iteraciones, empezando por tener los datos reducidos desde un principio, para tener un nuevo enfoque y poder realizar un estudio desde un punto de vista distinto.

Con la reducción se obtuvo un nuevo enfoque, ya que ahora tenemos bien agrupada la información para poder cotejarla. La siguiente imagen muestra un gráfico de los resultados del informe agrupados por empresas sin tener en cuenta los usuarios con las 100 empresas con más registros en un mes para la colección local. Estos se muestran en la imagen como la línea inferior de color negra y los datos de la línea verde son los registros mensuales del mismo mes para la colección de Analytics.

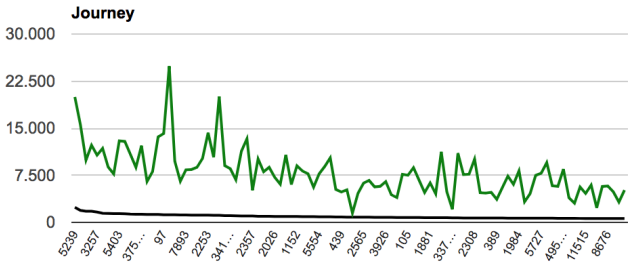

Fig. 2: Tabla de muestra de 100 empresas con el número de registros de visitas para cada una.

La siguiente figura muestran el índice de coincidencia que hay entre los informes de ambas colecciones, la línea azul muestra que índice hay por cada una de las empresas de la figura anterior, y la línea roja, es una media aritmética que se alimenta por cada empresa.

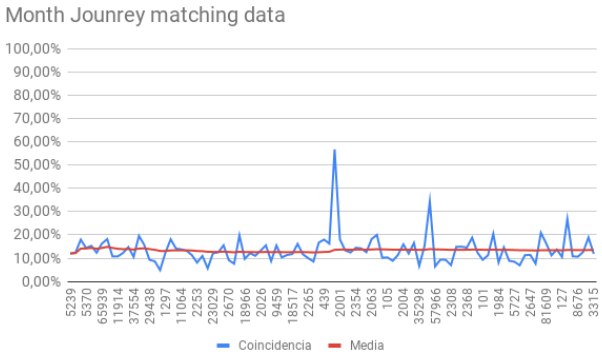

Fig. 3: Tabla de los índices de coincidendia de la Fig. 2.

Una vez realizado el informe, se procede a realizar la siguiente iteración, donde se analizaran estos datos y se empezará otro estudio para hallar patrones.

6 JOURNEY POR MÉS

El 'Journey' es el nombre con el que hemos atribuido a los informes de la anterior iteración, es debido a que los usuarios tienen agrupada la información por empresas de los registros, de esta forma tenemos una perspectiva mucho más amplia ya que ahora ya no enfocamos en ver registros sueltos, sino que sale la nueva directriz de preguntarnos, ¿cuántas visitas ha hecho un usuario en una empresa este mes? Con este nuevo modelo podemos comparar y ver que empresas son las más pobladas, compararlos con la de la otra colección, luego ver que usuarios tienen más datos referentes a esta empresa en este mes y finalmente comparar los registros de este usuario, ya que los usuarios con más distancia de cruce de una colección con la otra, nos aportan más información de lo que puede estar sucediendo en los datos.

Para empezar a ver las empresas con más visitas registradas a la web en el mes de la importación, se observó que la colección de Google Analytics tiene más del triple de datos que la colección local, es por esa razón de que es una buena candidata para someterla a estudio y ver que diferencias hay una de la otra.

Para realizar el estudio se analizó que usuarios tenían un mayor número de visitas en esta empresa en este mes y escoger entre los que tienen más visitas, además se escogió algún usuario en el punto medio de esta población por si también nos podía dar algún otro tipo de información.

Al escoger estos usuarios se analizaron los registros de estos en una colección y en la otra teniendo en cuenta la misma empresa seleccionada anteriormente. Con esta selección disponemos de información muy precisa de ambas colecciones con la que se puede cruzar la información, teniendo en cuenta las secciones de la página, sin la necesidad de tener en cuenta el tiempo, ya que con los datos tan reducidos se podía ver a simple vista los resultados.

El hecho de que se vea tan claro los resultados es porque la gran mayoría de los datos de la empresa de la colección de Analytics provenía de las fotos de la empresa, y no solo ocurría con un usuario, sino que en todos los usuarios sometidos a este estudio para esta empresa seguían el mismo patrón. El patrón se postulo con el título de 'Usuarios en la colección de Analytics que tienen gran numero de visitas en la sección de fotos de las empresas que no están siendo guardados por la colección Local'.

Una vez postulado el patrón anterior se observó que este patrón podía pasar la fase tres del procedimiento (comprobación de patrones), ya que para concluir que realmente este canon se cumple siempre, sé tiene que demostrar que podemos repetir el proceso de los usuarios para que una colección guarde nuestros datos en una colección y en la otra no. A la misma vez se tiene de hallar la sección en que estos inputs no actúan de la misma manera.

Al realizar la prueba se descubrió que el problema reside en el carrusel de fotos de las empresas, ya que la colección local solo guarda las trazas cuando una página es recargada, mientras que la colección de Analitics permite enviar una traza por imagen del carrusel vista. Como el carrusel se realiza con llamadas asíncronas al servidor, no es necesaria la recarga de página y por ese hecho la colección local pierde gran cantidad de visitas, ya que el gran numero de visitas en una empresa se hacen en la sección de fotos.

Con el punto anterior se cumple el procedimiento tres de la metodología de trabajo, antes de seguir con el punto cuatro, nos llego información de la persona encargada del SEO de la empresa informándonos de que en la colección de Analytics un usuario solo puede registrar 500 visitas por sesión. Debido a esta información se pospuso el proceso de la marginación del punto anterior para más adelante. Ya que antes se quiere realizar el procedimiento de estudio de patrones en el nuevo patrón hallado 'Usuarios de Analytics solo pueden hacer 500 hits en una sesión'.

Para estudiar este patrón se estudio si en los registros de Analytics, hay algún usuario con 500 hits, la respuesta fue que si, luego se observó si había algún otro que tuviese más de +500 y la respuesta fue que no, entonces con esto se halló un nuevo patrón que se debía comprobar que siempre se cumpliese. Tras realizar la comprobación, se observó que los usuarios solo podían llegar a realizar un máximo de 500 hits, se comprobó gracias a la herramienta de depuración, Google Analytics Debugger[4], y tal como se muestra a la imagen a continuación, cuando un usuario llega a 500 visitas en una sesión, la herramienta de Analytics deja de enviar información por seguridad.

Fig. 4: Respuesta de Google Analytics Debugger al exceder el máximo número de hits.

Una vez comprobado este patrón, se estudió que índice de la población de Analytics excedía el máximo número de hits por sesión permitidos. Al comprobarlo vimos que el ratio era tan pequeño que se podía despreciar su marginación de los datos, pero igualmente al cumplirse el patrón, se anotó como un factor que hace que los datos de ambas colecciones sean distintos.

Una vez finalizado con este patrón se procedió a realizar el punto cuatro del procedimiento, realizando un proceso de marginación para los datos que residen dentro del patrón de 'Usuarios en la colección de Analytics que tienen gran numero de visitas en la sección de fotos de las empresas que no están siendo guardados por la colección Local'.

Con este procedimiento se extrajo un nuevo informe con los datos reducidos sin tener en cuenta ambas fotos de las colecciones. Los resultados fueron sorprendentes, ya que la gran mayoría de los datos agrupados de cada colección se reduce su distancia del volumen de una respecto la otra. La siguiente figura muestra la misma gráfica del informe anterior, con la diferencia de que ahora no se tienen en cuenta las visitas a fotos de las empresas, dejando a ver la línea azul como los datos sin fotos de Analytics y la línea roja la de Local.

Para poder tener una visión más amplia de la figura anterior, se ha generado, en la figura siguiente, una nueva tabla para ver que porcentaje de coincidencia que hay en cada

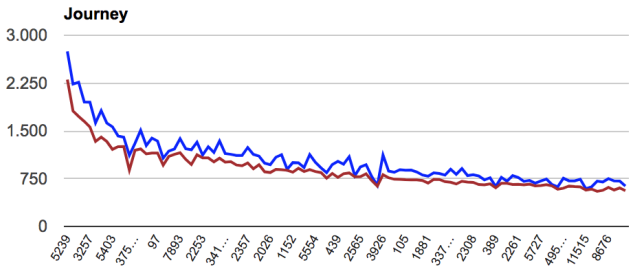

Fig. 5: Tabla de muestra de 100 empresas sin fotos con el número de registros de visitas para cada una.

empresa, que se muestra con una línea azul, y una media que se va alimentando por cada una de las empresas que se muestra con una liínea roja. Decir que se ha obtenido unos resultados con un dice mucho mayor, a cambio de reducir el volumen de datos de las colecciones.

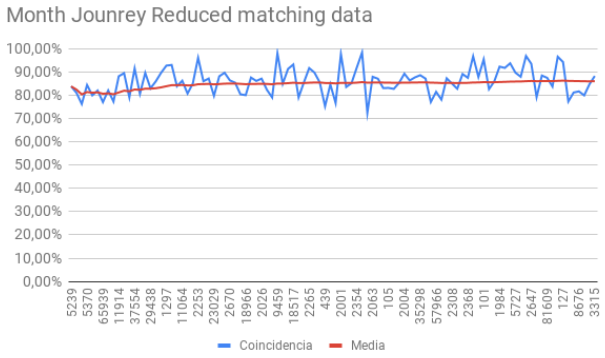

Fig. 6: Tabla de los índices de coincidendia de la Fig. 5.

Con este nuevo informe se discutió la idea de discriminar las fotos, ya que gran volumen de datos reside dentro de él. Al ser la sección menos sensible de las empresas, ya que respecto a las otras secciones es la menos importante, y debido al gran número de similitud que hay entre ambas colecciones sin tener constancia de las fotos, se decretó de que las fotos no se tuvieran en cuenta para tener una correcta convivencia entre ambas colecciones. De esta forma se normalizó que no se tuviera en cuenta las fotos en la migración final de los datos.

Finalmente, se concluyó en esta iteración de que el índice de coincidencia del Journey de cada colección era muy parecido a la de la otra colección, se dio por satisfecho los resultados y a partir de este punto ya se tiene el Ok del proyecto para realizar la migración final de los datos y con esto el cumplimiento de todos los objetivos del proyecto menos el segundo (ver que conflictos residen en cada una de las colecciones respecto la otra).

Ya que seguimos desconociendo que factores intervienen, para hacer que ambas colecciones no sean idénticas, por esta razón el proyecto no se finalizó en este ciclo, sino que se decidió continuar para investigar otros factores que hacen incoherentes los datos de ambas colecciones.

7 Experimento 2.0

El experimento 2.0 es el reemprendida del experimento del primer ciclo, pero ahora con la diferencia de que tenemos el Journey mensual con los datos conflictivos marginados. Con esta iteración se quieren hacer otra vez el cruce de datos, pero solo en esos usuarios que no tienen el mismo volumen de registros en ambas colecciones. En esta ocasión, orientar de forma correcta, sin realizar modificaciones o acciones que deriven a hacer inconsistentes los datos, y poder analizar esos datos a los que no se pueden cruzar, para estudiar si existe algún patrón en ellos, para posteriormente marginarlos.

Para realizar este proceso hemos seleccionado los candidatos a través del Journey mensual, ya que de esta forma, gracias a la comparativa de ambas colecciones, podemos seleccionar un usuario con diferentes números de visitas a una empresa y analizar que registros hay diferentes en él.

ID_USER	ID_EMPRESA	MONTHS	YEARS	NUM_PAGE_VIEWS	NUM_PAGE_VIEWS_WITHOUT_PHOTOS	NUM_PAGE_VIEWS_LOCAL	NUM_PAGE_VIEWS_WITHOUT_PHO...
77661	26694	3	2018	19	18	14	13

Fig. 7: Muestra de registro candidato a analizar del informe del Journey mensual.

La imagen anterior, muestra un registro del informe mensual, donde se muestra que un usuario tiene registrados 19 registros en Analytics y 14 en la Local, y con la reducción esta pasa a ser de 18 y 13 respectivamente. Este es un buen candidato a someterse a análisis debido a que este registro es anómalo, ya que con el proceso de marginación de fotos sigue teniendo la misma variación, eso quiere decir, que este candidato cumple otro patrón que hace que ambos números no sean el mismo.

Por esa razón hemos seleccionado esas empresas con el índice de coincidencia más bajo y ver que usuarios tienen mayor variación en el número de visitas en la misma empresa, para posteriormente ver si existe alguna norma que cumple que esos valores de los registros no sean los mismos.

Al hacer la búsqueda lo primero que vimos es que gran cantidad de los usuarios de la colección Local no se tienen constancia en la colección Analytics, y pasa lo mismo a la inversa, hay usuarios en Analytics que no están registrados en la Local. Para el total de datos del informe, el 12.47 % de los usuarios tienen información en la base de datos local y no en Analytics, y el 8.01 % para los registros que tienen información en Analytics y no en la colección Local.

Al no tener una noción clara de cuál podría ser la causa, se realizó una búsqueda por distintos portales web donde usuarios con la misma problemática hablaran de esta. Lo más sorprendente es que hay gran cantidad de información acerca de la manca de usuarios que no registran su navegación en la base de Google Analytics.

Entre los hilos sobre este tema más recurrente se hace referencia de esos ADD-ONs que son capaces de bloquear el tráfico hacia terceros, como uBlock[5]. o en su día AdBlock[6], que hacen que se envíen los mensajes con las trazas a Universal Analytics, uno de los debates que nos ayudo más fue el de Markus Rene en el foro de freecodecamp[7] que nos da información sobre la conectividad de los usuarios, que nos permite encontrar nuevos patrones que no se tenía constancia.

> ”I was going nuts on this for quite some time now and for me the reason was quite simple: uBlock was blocking Google Analytics. Since with the new firefox you have your addons on android/mobile version too now, it took me a while to get behind this. I am posting this here for others because Google Search led me here as one of the first search results.”[7].

La cita anterior, es un extracto del mensaje de Markus Rene en el debate de Freecodecamp, este texto nos ayudo mucho con el progreso, ya que tal y como lo interpretamos, nos da información sobre que puede haber agentes que bloqueen los mensajes que se envían de las trazas. Por esa razón, gracias a la ayuda de este post se halló un patrón que hace que un porcentaje de información, que la colección de Analytics debería recoger, no lo haga debido a estos agentes. Después de descubrir esta problemática se quiso saber que porcentaje aproximado es el que se llega a perder, al ser varios agentes los que intervienen y son atemporales, no se sabe en que momento han podido intervenir. Por lo que este patrón se le ha etiquetado como patrón temporal. Ya que no su estudio no puede hacerse de un método convencional y se requiere gran cantidad de recursos para estudiarlos, es por esa razón que se realizó una búsqueda entre distintas fuentes para saber cual podría ser la magnitud.

Se encontraron muchos agentes temporales que en determinadas ocasiones han interrumpido el paso de mensajes de Google Analytics. Uno de los patrones más frecuentes, es el uso de ADD-ONs que utilizan los usuarios para bloquear tráfico de dominios de terceros como Adblock. Ya que en distintas ocasiones el mismo agente ha permitido o no enviar información, es por eso que es imposible determinar cuando intervinieron esos agentes. La única información que se ha podido sacar, es un articulo de una página, donde se habla de este mismo problema[8].

En el artículo anterior no se realiza un estudio exhaustivo en el que se demuestre las estadísticas mencionada. Pero si es cierto, que esos agentes no permiten el paso de información a Analytics.

Llegados a este punto se asume que el 12.47 % de los datos que no están en Analytics son debidos a patrones temporales que interrumpen el guardado de estos usuarios, ya que a diferencia de la colección Local, esta no se ve interrumpida debido a que no forma parte de un dominio de terceros.

Una vez finalizado con el estudio de Analytics, nos queda un 8.01 % de datos que la colección Local no tiene constancia. Debido a que los datos de esta colección no están localizados en un dominio de terceros, no se ven afectados por agentes que hagan aparecer patrones temporales y, por lo tanto, deberían constar todos los datos.

Al no ser así se analizó los posibles motivos que hacían que el 8.01 % de los datos no estuvieran registrados. Para este punto se analizó como se lanzan los eventos en la página para intentar conseguir algo de información que nos permitiera descubrir si existía alguna coincidencia.

Se descubrió que los eventos funcionan de la siguiente manera. Primero se ejecuta el evento que está en el head de la página, por lo que es lo que se ejecuta al inicio de la carga y, al final de la carga de la página, se ejecuta el script de la herramienta local para guardar datos.

Con el conocimiento del orden de los eventos, se sabe que la base de datos local no registra los datos si la carga de la página no ha finalizado. Por ejemplo si un usuario va a una empresa y antes de finalizar la carga ve en el menú que puede ir a otra sección y no espera a que finalice la carga , solo enviará la traza Analytics y la base de datos local quedará sin registrar esta información.

ID	ID_USER	ID_USER_LINKED	ID_ITEM	ID_TYPE	ID_ACTION	REDUCED	START_TIME
158638544	1078318	NULL	5153	1	1	/vendors/item/real_weddings	2018-06-01 12:13:54
158638568	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:14:00
158638642	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:14:30
158638678	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:14:40
158638934	1078318	NULL	5153	1	1	/vendors/item/real_weddings	2018-06-01 12:16:11
158638994	1078318	NULL	5153	1	1	/vendors/item/faqs	2018-06-01 12:16:19
158639018	1078318	NULL	5153	1	1	/vendors/item/real_weddings	2018-06-01 12:16:24
158639048	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:16:30
158639060	1078318	NULL	5153	1	1	/vendors/item/real_weddings	2018-06-01 12:16:35
158641076	1078318	NULL	5153	1	1	/vendors/item/real_weddings	2018-06-01 12:24:40
158641752	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:28:21
158641806	1078318	NULL	5153	1	1	/vendors/item/real_weddings	2018-06-01 12:28:41
158641858	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:28:53
158641902	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:29:04
158641966	1078318	NULL	5153	1	1	/vendors/item/real_weddings/item	2018-06-01 12:29:26

Fig. 8: Muestra de registros realizados por un usuario de demostración.

En la figura anterior se muestran los registros de usuario de demostración, hecho por nosotros, visitando una empresa por orden cronológico, donde este usuario ha sido capaz de navegar entre ítems de la empresa sin tener que visitar la lista de los ítems intermedios. Hay dos opciones para realizar esto, la primera, es que habrá con diferentes pestañas los ítems, y la segunda y es la cierta, es que el usuario al navegar por un ítem y cuando quiera cambiar a otro, lo haga con suficiente rapidez, sin tener que ejecutar el script de la herramienta Local.

Se ha comprobado que todos y cada uno de los patrones se cumplen, pero es muy difícil de identificar que registros de las colecciones cumplen estos patrones, ya que como hemos visto no es tarea trivial identificarlos. Algunos se desconoce en que momento han aparecido y cuando lo han dejado de hacer, y otros patrones se sabe de ellos pero no se pueden distinguir de entre los demás registros como es el caso de la navegación rápida del usuario. Por lo que hace una tarea muy difícil, marginar dentro de las colecciones estos registros.

Debido a que el índice de coincidencia, después de realizar el proceso de marginación de las fotos, se considera como aceptable la combinación de ambas colecciones, por lo que se decide a no tenerlos en cuenta para realizar otra migración. De todas formas, para cumplir con los objetivos se quiere saber que eventos hacen que los datos no sean exactos, y debido al ser patrones que se han demostrado que cumplen que haya datos localizados en una colección y no en la otra, se tienen de tener constancia de ellos, aun sin tenerlos en cuenta para la migración de los datos.

Al final de este ciclo, se decide preparar la importación del Journey diario, este paso es el último antes de analizar los datos finales de la migración. Trata de separar el Journey de los usuarios en días en vez de meses, de esta forma podemos tener los datos en un formato más adaptado para poder analizar a los usuarios por día y poder así analizar su comportamiento a diario.

Este proceso se ejecutará en la siguiente iteración, debido a que antes de ejecutarlo se quiere revisar que se tenga en cuenta todos los aspectos del proyecto y, se quiere hacer un análisis de todos los patrones encontrados para ver cuales de ellos se usarán en la importación final.

Se quiere hacer un recordatorio que este artículo no ataca

al estudio que se hará posteriormente a estos usuarios, su comportamiento y análisis recae en el dominio de otro proyecto. En este solo se estudian los datos para que se tenga un formato correcto, se tenga una visión global de todos los datos y saber cuál es su lugar en este entorno, ya que como hemos visto no siempre es igual para ambas colecciones.

8 JOURNEY POR DIA

Esta sección consiste en la última fase del proyecto, donde se detalla el conjunto de patrones tenidos en cuenta para realizar la migración de los datos. Esta migración se quiere en un formato concreto ya que va a servir para otro proyecto de análisis de comportamiento de usuarios. Se quieren importar ambas colecciones a la colección Local en formato de Journey Users diario, por lo que queremos separar por días, usuarios y empresas, y realizar un informe completo de ambas colecciones, una vez realizado se quiere combinar ambos informes en uno, manteniendo los resultados de cada colección en columnas separadas.

También se quiere un formato óptimo, ya que no es posible migrar todos los registros de la colección de Analytics a la otra. Por esa razón se hace una reducción de los datos en informes diarios, ya que cumple con el formato que se quiere al final y un grado de optimización que haga que el proyecto en menos de un día pueda estar importado. Finalmente, al tener los informes segmentados por periodos de tiempo, en caso de que el proceso se interrumpiera, se podría retomar des del último periodo de importación, lo que hace el sistema tolerable a caídas.

La importación que se quiere realizar es simple, se tiene que realizar de la misma manera que se realizó con el Journey de un mes, pero en vez de agruparlos por mes se tiene que hacer por día. Al tener un volumen mucho mayor de datos hace que la migración no sea igual de sencilla, por lo que los tiempos de ejecución y de gestión de los datos se disparan de forma exponencial, es decir que se tarde alrededor de 12 horas a realizar la importación. Debido a que es un servicio externo a la empresa y hacer uso de él tiene un costo económico, se idea una forma distinta para no estar las 12h con el servicio activo.

Por esa razón se decide a hacer la importación por semanas y gestionando en el entorno PHP para poderlo separar por días, de esta forma el proceso es mucho más óptimo, ya que ejecutar un proceso de BigQuery por un día es un malgasto de recursos innecesario. Pero para 1 semana es aceptable y, la gestión de los datos para el servidor PHP no es muy costoso, si fuera mayor el volumen de datos sería mucho más costoso.

Haciendo la modificación del proceso, el tiempo final de importación es un poco menos de 7h, lo que quiere decir que casi se reduce a la mitad el tiempo de ejecución, y se hacen 7 veces menos consultas a Analytics para realizar la importación. Lo que es un éxito, ya que para la empresa es una reducción de los costes económicos y del tiempo muy importante.

Para realizar la migración de los datos se tienen en cuenta los siguientes criterios y patrones:

- La franja horaria de cada una de las colecciones
- Las fotos no deben constar en ninguna de las colecciones
- Se debe tener en cuenta el cambio de horario de verano/invierno para las franjas horarias.
- Debe realizar-se por semanas y separlas luego por días.
- Las inserciones de datos tienen que realizarse por bloques.
- En caso de caída se debe retomar des de la última semana importada correctamente.
- Se debe hacer la importación de modo incremental, partiendo del primer día hacia delante.

Con estos criterios podemos realizar la importación final, los puntos más importantes son: tener en cuenta el patrón de las imágenes y la franja horaria, ya que sino muchos de los datos del informe de un día podrían verse reflejados a otro día, por lo que daría paso al error, ya que al ser días distintos no se pueden cruzar los datos de coincidencia.

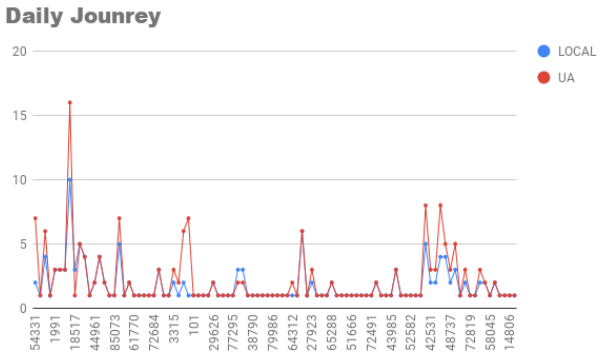

Fig. 9: Tabla de muestra de 100 empresas del Journey diario con el número de registros de visitas.

En la figura anterior se expone una muestra de 100 items de la población del journey por día de la colección de Analytics y de la Local, líneas roja y azul respectivamente. En la siguiente figura se muestra una tabla con el índice de coincidencia de cada empresa y la media por cada una de ellas con referencia a la figura anterior.

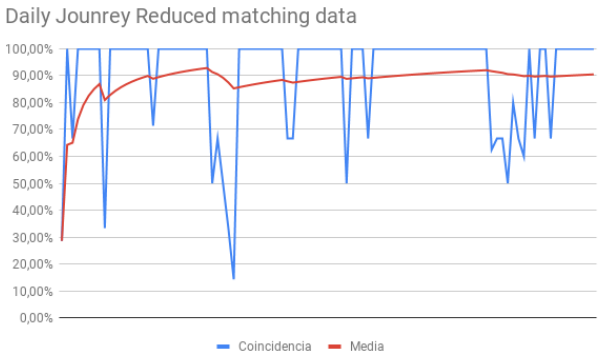

Fig. 10: Tabla de los indices de coincidendia de la Fig. 9.

Al finalizar este proceso de importación, se cumplen finalmente todos los objetivos del proyecto. Para poder finalizar el proyecto y cerrarlo con éxito se pasa a realizar un estudio final de los resultados obtenidos. Este estudio es una visión de principio a fin del proyecto y que hemos obtenido

tras realizarlo. El estudio se expone en la siguiente sección 'Resultados'.

9 Resultados

En esta sección se estudian y se exponen los resultados de los datos obtenidos tras la migración realizada en la sección anterior. Se quiere ver desde que punto parten ambas colecciones al inicio del proyecto, y hasta donde han llegado tras la migración.

Al principio de este proyecto, vimos que realizar el estudio de las colecciones no sería una tarea fácil, ya que partiamos de un punto que los datos aparentemente no tenían nada que ver unos con otros.

Se partía desde un punto que ni el 2 % de los datos coincidian, debido a que se desconocía que las colecciones estaban guardadas en distintas franjas horarias.

Al final de este proyecto, hemos llegado a cruzar el 69.5 % de los registros del informe del Journey diraio. Eso quiere decir que los registros de usuario, empresa y día del informe de cada una de las colecciones se han podido emparejar en un 69.5 %. El otro 30.5 % restante son esos registros que complementan la información total que hemos obtenido, siendo el 23.7 % de Google Analytics y el otro 6.8 % la base de datos Local. Estos registros de este informe que no se han podido cruzar, son los que salen de los patrones que no hemos podido marginar debido a su complejidad, pero aun así sabemos que no son errores, sino datos que la otra colección no ha podido registrar.

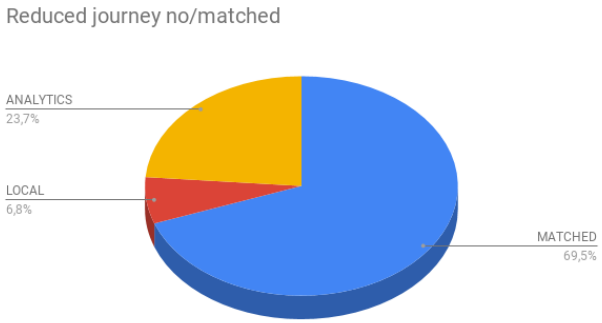

Fig. 11: Gráfico de porcentajes de coincidencia de los registros tras la importación.

Después de analizar los informes a nivel de registro, se ha hecho lo mismo a nivel de valores para saber exactamente cuantos datos hemos podido alcanzar a emparejar.

Los resultados obtenidos tras el análisis fueron que el 42 % de los datos totales de ambas colecciones se han combinado. Eso qes un buen indicio de que el proyecto ha servido para aumentar del 2 % al 42 % el índice de coincidencia en los datos. Este aumento no se ha debido a una modificación de los datos, sino a que se ha realizado una marginación de datos y una adaptación en el horario para que no hubiese errores. Tras realizar este proyecto, sabemos que el valor inicial era tan bajo debido a que el carrusel de fotos no guardava información en la colección local, haciendo que los datos de la colección de Analytics se disparasen hasta un punto que los datos de la colección Local se viesen reducidos hasta quedar ocultos tras el gran volumen de la otra colección.

Tras analizar a fondo los datos, se sabe que el 35.8 % de los datos de Analytics se han añadido dentro del informe del Journey diario y el 22.2 % de la colección Local.

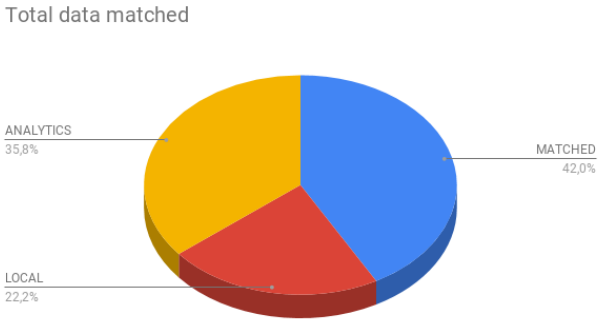

Fig. 12: Gráfico de porcentajes de coincidencia de los valores de los datos tras la importación.

Al acabar con este análisis, se decidió dar por válidos todos los datos de la migración realizada. Al tener un nuevo conjunto de datos mayor al que teníamos, se querie ver, si puede solventar una problemática recurrente en la página. El problema es que hay empresas que sin tener conocimiento de que usuarios las visitan, esos usuarios son capaces de realizar acciones que tenemos registradas. Se sabe de donde pueden salir esas acciones, aunque no se sabe si todas las acciones vienen del mismo punto del que tenemos constancia.

Tras realizar la importación se quiere observar si gracias al añadir el conjunto de datos de la colección de Analytics, somos capaces de ver si para estos datos con acciones pero sin visitas tras la importación siguien teniendo el mismo número de 107290. Al analizarlo se observó que, para estos registros de los cuales no teniamos constancia de visitas, con el agregado de la otra colección 13270 de los registros han conseguido visitas.

La figura de a continuación se muestra como hay datos sin registro de Visited, que son de la colección Local, se genera un guardado, se solicita un presupuesto o se contrata sin tener constancia de que el usuario haya visitado esa empresa. Tras la importación podemos ver en la última columna, Visited_UA, que son los datos de Google Analytics, y hay información de visita para esos usuarios que no teníamos constancia de visita anteriormente. Eso quiere decir, que realmente han habido usuarios que han visitado una empresa se han generado acciones que se han guardado, pero no se ha tenido constancia de la visita del usuario desde la colección Local, en cambio en la colección de Analytics si que se han guardado, por esa razon tras la importación podemos observar que estos usuarios ahora tienen guardadas sus visitas.

Finalmente, se concluye este proyecto con la aprobación del equipo de Machine Lerning[9] de la empresa. Ya que ellos utilizarán esta información que hemos migrado para poder alimentar sus estadísticas, y poder realizar esta nueva información como una herramienta más para el desarrollo de sus tareas.

ID_USER	ID_EMPRESA	VISITED	SAVED	LEAD	BOOKED	FECHA	VISITED_UA
5939	13933	0	1	1	0	2017-05-17	2
5939	28188	0	1	1	0	2017-05-17	2
5939	29824	0	1	1	0	2017-05-17	1
5939	64767	0	1	1	0	2017-05-17	2
14063	23418	0	1	1	0	2017-05-24	1
23375	392	0	1	0	0	2017-05-15	1
23375	5077	0	1	0	0	2017-05-13	1
36335	34561	0	1	0	0	2017-05-13	1
38808	35527	0	1	1	0	2017-05-05	2
38808	35667	0	1	1	0	2017-05-05	2
38808	36600	0	1	1	0	2017-05-06	1
38808	45483	0	1	1	0	2017-05-05	2
49058	27285	0	1	1	0	2017-05-07	2

Fig. 13: Muestra de registros sin visitas locales, con acciones y visitas de Analytics.

10 Conclusiones

Con este artículo hemos podido analizar un gran conjunto de datos enfocados en un entorno web, se ha realizado una comparación entre dos bases de datos distintas que aparentemente deberían de tener los mismos registros, tras realizar el análisis se ha visto que no era así, ya que mucha cantidad de una colección no estaba localizada en la otra. Tras realizar un estudio se concluyó que gran cantidad de la información que falta de una de las colecciones es debido a que una de las secciones de la página no registra datos para una de las herramientas, en concreto la que almacena los datos a la colección Local. Tras realizar este descubrimiento, se vio que seguían faltando algunos datos de ambas colecciones, puesto que las colecciones seguían siendo diferentes. Se realizó otro estudio para ver cuales podían ser las causas y, se observó que las causas no se pueden marginar dentro de las colecciones, debido a su difícil identificación. Puesto a que no se puede hacer la migración final teniendo en cuenta todos los patrones, se realiza la importación marginando los datos de la sección que solo graba información de una de las colecciones, y finalmente se hace una comparativa entre ellas para ver los resultados obtenidos de este proyecto.

Agradecimientos

En esta sección quiero dar reconocimiento y las gracias a bodas.net por dejarme colaborar con ellos para la realización de este proyecto, el equipo de desarrollo Bravo que me han prestado recursos para la realización de este proyecto y en especial a David Guillament y Yujiro Koyama. También a la Universidad Autónoma de Barcelona (UAB) por hacer un seguimiento y evaluación del proyecto y finalmente a M. Fauqued por la corrección de este artículo

Referencias

[1] Clifton, B. (2012). Advanced Web metrics with Google Analytics. Hoboken, N.J.: J. Wiley Sons, Inc.

[2] Google Cloud BigQuery. (2018). BigQuery - Analytics Data Warehouse — Google Cloud. [online] Available at: https://cloud.google.com/bigquery/ [Accessed 1 Jul. 2018].

[3] Sheldon, R. and Moes, G. (2005). Beginning MySQL. Indianapolis, IN: Wiley Pub.

[4] Google Analytics Debugger. (2018). keithclark/gadebugger. [online] Available at: https://github.com/keithclark/gadebugger [Accessed 1 Jul. 2018].

[5] UBlock. (2018). gorhill/uBlock. [online] Available at: https://github.com/gorhill/uBlock [Accessed 1 Jul. 2018].

[6] Adblock. (2018). AdBlock. [online] Available at: https://getadblock.com/ [Accessed 1 Jul. 2018].

[7] freeCodeCamp. (2018). How to prevent your analytics data from being blocked by ad blockers. [online] Available at: https://medium.freecodecamp.org/save-your-analytics-from-content-blockers-7ee08c6ec7ee [Accessed 1 Jul. 2018].

[8] Quantable. (2018). How Many Users Block Google Analytics, Measured in Google Analytics - Quantable. [online] Available at: https://www.quantable.com/analytics/how-many-users-block-google-analytics/ [Accessed 1 Jul. 2018].

[9] Sugiyama, M. (2016). Introduction to statistical machine learning. Waltham, MA: Morgan Kaufmann Publishers.