View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by El Servicio de Difusión de la Creación Intelectual

Un caso de big data punta a punta: análisis de datos de transporte y su uso en el negocio.

Melani Camilo, Echagüe Juan V., Torre Zaffaroni Joaquín, Yankelevich Daniel

cmelani@pragmaconsultores.com

Keywords: Big Data, SUBE, Análisis de la Demanda, Analytics, Hadoop

Introducción.

En este artículo se presentan los resultados de un proyecto de análisis de datos de una empresa de transporte, que involucró la recolección, preparación, visualización, transformación y análisis de 3 años de datos de viajes de colectivos, incluyendo boletos y posicionamiento geográfico. Este caso cubre el proyecto de punta a punta, incluyendo la incorporación de los resultados en el proceso de negocio.

Nosotros sostenemos que una característica clave de los proyectos de big data debe encontrarse en el proceso que se lleva a cabo y que inicia con la captura de grandes cantidades de datos, pasando por el procesamiento (que en muchos casos requiere una infraestructura especial o particular, con más de una computadora, en modo distribuido) hasta el análisis y el aprovechamiento de la información en el negocio. En nuestro punto de vista, este último paso (la inserción de la información en la toma de decisiones) es tan importante como el uso de bases NoSQL o Hadoop o procesar varios terabytes.

Datos.

SUBE es una tarjeta prepaga emitida por el Gobierno Nacional argentino para facilitar la movilidad en el área metropolitana. Puede usarse en los medios de transporte públicos en la región Metropolitana de Buenos Aires y el interior del país. La red de uso está compuesta por 11.000 colectivos, 5 líneas de subtes y las líneas ferroviarias metropolitanas, y diariamente vende 12 MM de boletos de transporte.

Los datos para este análisis fueron provistos por una empresa de trasporte de mediano tamaño del conurbano bonaerense. Cuenta con 110 colectivos, 3 líneas, 11 ramales que diariamente recibe las transacciones realizadas. Esta empresa cuenta con un servicio que de forma on-line informa la posición GPS de cada colectivo. Accedimos a 3 años de venta de boletos (40 millones) y la posición de cada colectivo (150 millones).

Infraestructura y Metodología de Trabajo

Para realizar el procesamiento de los datos es necesario contar con capacidad de almacenamiento, acceso a la información y poder de cálculo adecuados. Para este caso utilizamos infraestructura propia con tecnología Apache HDFS [1], Hive [2], R [3] y Hadoop [4] sobre una estructura de 6 nodos.

La metodología de trabajo fue exploratoria, pero seguimos un esquema de trabajo propio en el cual la identificación de criterios de evaluación del negocio formó parte del proyecto desde un inicio [5].

Limpieza, comprensión y análisis de datos.

La preparación de datos es una parte importante en un proyecto de big data [6], de hecho en muchos casos el "data cleansing" y preparación inicial toma más tiempo que el análisis. En este proyecto, la preparación de datos incluyó identificar y subsanar varias limitaciones de los datos, por ejemplo, los relojes de los lectores del sistema SUBE y los GPS no están sincronizados. Las granularidades de las diferentes fuentes de datos no es la misma, en el caso de los GPS las posiciones se reportan en cada minuto. Asimismo, el trabajo se realizó sobre datos anónimos lo que requirió trabajo adicional.

Las tareas de análisis incluyeron la elaboración de histogramas, gráficos de series temporales, heatmaps en varias variables, generación de imágenes geo localizadas de la concentración de venta de boletos, identificar los trayectos de mayor demanda, relacionar los pasajeros frecuentes con el tiempo entre trayectos y generación de grafos.

Gran parte del análisis se focalizó en identificar casos o preguntas del negocio: qué era lo que el negocio consideraba interesante para conocer y a qué le otorgaba valor, identificar el comportamiento de los clientes que permitieran su segmentación. Contar con toda la serie histórica desde que se implementó la tarjeta SUBE en esta empresa, nos permitió observar con sumo detenimiento la curva de adopción del sistema y el comportamiento de reemplazo del modelo anterior. Este mecanismo permite analizar y establecer patrones sobre el proceso de adopción de políticas públicas. Encaramos un estudio multiescala sobre la densidad de venta de boletos en diferentes horarios (ver Ilustración 1). Los primeros datos obvios se reflejan claramente en la combinación de datos georreferenciados y clustering, y se observa como en horarios matinales, las personas se desplazan desde barrios periféricos a lugares de concentración comercial o industriales, y por las tardes este proceso se revierte.

Ilustración 1.Clustering y heatmap de la posición de la venta de boletos.

Conclusiones.

El dataset contiene información muy prometedora y en ese sentido las expectativas eran muy altas. A la vez, el objetivo no era realizar análisis sociológicos o generales sobre los datos, sino lograr información relevante para el cliente, en particular información accionable que permitiera mejorar su posición de negocio o responder preguntas de negocio. A modo de ejemplo, el uso repetido y consecutivo de la misma tarjeta en un lapso reducido. Esto puede tener relación directa con un abuso de las políticas sociales implementadas, ya que ciertas tarjetas tienen un cuadro tarifario diferente, asociado a la situación social de cada

persona. Sin embargo, descubrir, estudiar o resolver esto no es objetivo de nuestro análisis salvo en lo que se refiere a la información relevante para la empresa.

Este trabajo de análisis de datos permitió al cliente contar con herramientas para conocer de forma profunda y con altísimo nivel de detalle la distribución de la demanda. Preguntas clave del cliente como "patrones de venta durante el dia" (ver Ilustración 2) o "cuál es el patrón de viajes de los clientes frecuentes" pudo responderse con información precisa, así como descubrir patrones inesperados.

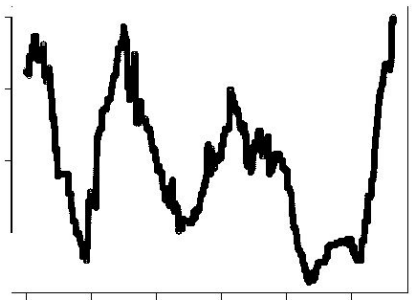

Ilustración 2. Crosscorrelacion entre secuencia de venta y venta del día para un colectivo. Se observan en los picos las 4 veces que se inicia la vuelta.

El piloto realizado es la primera etapa de un verdadero estudio en profundidad de los datos, que el cliente debe aprobar y suscribir. En algunos casos se utilizaron herramientas de graficación para presentar los datos al cliente en forma intuitiva, y algunos resultados preliminares de correlación de datos. En la segunda etapa este análisis podría llevarse a cabo con mayor detalle.

Este trabajo permitió agregar valor a la empresa mediante varios mecanismos, ya que conocer el detalle de la demanda habilita el uso de herramientas comerciales en forma sistemática e informada, que de otra forma se aproximan por la intuición o la experiencia. La intuición no siempre coincide con la situación real y actual en la dinámica del negocio, ya que refleja el conocimiento de muchos años y una visión en algunos casos subjetiva de una realidad cambiante. Para poder mejorar hay que saber medir y en este caso se logró responder a las preguntas de negocio con el soporte de datos y con alta definición.

Referencias.

[1] Apache HDFS , «Available at http://hadoop.apache.org/hdfs,» [En línea].

[2] Apache Hive, «Available at http://hive.apache.org,» [En línea].

[3] R Core Team, R: A language and environment for statistical computing., Vienna, Austria.: ISBN 3-900051-07-0, URL http://www.R-project.org/, 2013.

[4] Apache Hadoop, «Available at http://hadoop.apache.org,» [En línea].

[5] Pragma consultores, «4 Pasos para ir del dato a la decisión,» 12 2014. [En línea]. Available: http://www.pragmaconsultores.com/ar/SiteCollectionImages/Revista/.

[6] V. Cotik, P. Lujan, D. Scotton y D. Yankelevich, «A Swiss army knife approach to DQ assessments,» *IJIQ,* vol. 1, nº 2, pp. 145--161, 2007.