DATA SCIENCE APLICADA AL DESARROLLO DE PRODUCTOS

Quizás algunas de las referencias que se mencionan en este artículo sólo sean memorables para los que nos peinamos con canas blancas. Desde ya le pido disculpas a las jóvenes mentes que lean este texto. Aun así deseo aportar a la conversación sobre esta interesante historia que incluye integración de los *data scientists*, utilización de datos abiertos y la evolución de plataformas tecnológicas, entre otr os entremeces.

PABLO GARCÍA

Gerente de Investigación y Arquitecto Jefe de Telefónica Investigación y Desarrollo de Chile. Anteriormente se desempeñó como arquitecto principal en Microsoft durante más de una década y anteriormente como líder técnico y director de la empresa Urudata. Es profesor de Programación, Estructuras de Datos y Algoritmos para el Programa de Desarrollo de las Ciencias Básicas de Uruguay. Es Máster en Computación y sus áreas de interés principales son la automatización de procesos y el procesamiento de secuencias planas de ADN.

pablo.garciab@telefonica.com / @pc_garcia

Hace algunos años vi una presentación de Jim Gray –un científico que la mayoría de la gente recuerda por sus trabajos en transacciones distribuidas o por el ACID test–, donde trataba de difundir un concepto que había expuesto en la Academia de Ciencias Americana, en un discurso ante el Congreso. En esa charla, él hablaba que la ciencia había pasado por varios paradigmas.

El primero, que supuso un nuevo mecanismo de avance sustancial para la ciencia, lo estableció Galileo con el establecimiento de la ciencia experimental. El segundo es la de la ciencia teórica. Ésta aparece con el trabajo de Newton sobre los planetas, el que establece que en su movimiento de traslación estos recorrían áreas iguales en tiempos iguales.

Luego, la teoría atómica nos ayudó a establecer el paradigma de las simulaciones. En ello primero se teorizaba, luego se construía una simulación y posteriormente se la contrastaba con el experimento. Finalmente, ya en nuestros días, llegamos a un nuevo paradigma, caracterizado por el volumen enorme de datos. Fuentes como el secuenciamiento de ADN, la ciencia del clima y la telemetría de muchísimos dispositivos conectados, hace que la aplicación de técnicas computacionales de descubrimiento de patrones, establezca un nuevo paradigma para el avance de la ciencia. A este nuevo paradigma podemos llamarle la Ciencia de los Datos o en su acepción anglosajona: *Data Science.*

Ese nuevo paradigma ya traspasó las fronteras de la ciencia y hoy vemos una enorme curiosidad dentro de las empresas por mutar su estrategia de *datawarehousing* a un modelo que los aleje de los costosísimos procesos de Extracción, Transformación y Carga, eliminando la Transformación y cargando los datos en bruto, aplicando luego fuerza de procesamiento para transformarlos al momento de analizarlos.

Como todo en la vida, este nuevo paradigma tiene ventajas y desventajas. La principal utilidad es que la tecnología que se usa para estos procesos, en el día de hoy, es un *commodity*. También se ha convertido en un *commodity* el tener enormes capacidades de almacenamiento y de procesamiento *on-demand,* a precios ridículamente bajos, aunque esto es la parte menos importante de la ecuación.

Lo que hoy complica es que el equipo de *Business Intelligence* (BI) no puede hacer la transición de ordenador de la información a *data scientist*, siendo éste el dilema en que se encuentran hoy las organizaciones: todos miran los reportes de las consultoras como Gartner, y muchos desean abordar este tema dentro de su organización, pero se encuentran con que los *data scientists* son pocos y a la mayoría no les gusta hacer reportes sofisticados de marketing.

DATA SCIENTISTS INTEGRADOS AL NEGOCIO

Para aplicar *Data Science* hay que dejar claro que esto implica muchos requerimientos, pero podemos establecer algunas bases mínimas, tales como: saber procesar grandes volúmenes de datos,

saber analizarlos en forma rigurosa, comprender la relevancia estadística de los resultados obtenidos y saber del negocio para poder interpretar los resultados.

Hace unos años atrás hice un Magíster en Bioinformática, y cuando tomé las materias electivas del área biológica no me pareció relevante tomar Evolución. Sin embargo, al avanzar en la maestría, me pasó que los biólogos descartaban rápidamente resultados aplicando la teoría de la evolución con un razonamiento que resultaba muy difícil de formalizar. Era su conocimiento de cómo cambian organismos y poblaciones, influenciados por la evolución, lo que les permitía rápidamente descartar un grupo de resultados que parecían matemáticamente relevantes, y concentrarse en los que tenían sentido biológico.

Me refiero con esto a que, para poder aplicar técnicas y conocimiento de estadísticas y *machine learning* en el procesamiento de bases de datos de un negocio y entregar algo útil como salida, se necesita conocer del negocio en particular. No basta simplemente con traer un experto y que él realice el reporte, es necesario integrarlo a la organización y empaparlo en el negocio para que aparezcan los mejores resultados.

Podemos tentarnos a armar un equipo multidisciplinario, pero con esto no alcanza. Por ello, de las cosas que estamos realizando en el Centro de Excelencia Internacional de Investigación y Desarrollo de Telefónica (en alianza con Corfo y la Universidad del Desarrollo) además de integrar equipos de *data scientists* con especialistas en minería, agricultura de precisión y ciudades inteligentes, es a no escatimar recursos para que nuestros *data scientists* aprendan en profundidad el contexto en el que están desarrollando su trabajo.

Existen enormes oportunidades de hacer minería de procesos y aplicar diversas técnicas de minería de datos, pero para hacer la diferencia profunda dentro de la organización se necesitan recursos humanos, tiempo para dominar las técnicas y madurar el conocimiento del negocio. Todavía estamos en etapas muy iniciales, lejos de recoger los enormes beneficios que promete el estado del arte de la tecnología.

LOS DESAFÍOS EN DATA SCIENCES

Hay muchísimo por hacer en la tarea de cimentar la importancia de la aplicación de data science en el desarrollo tecnológico. En Telefónica I+D Chile encontramos a diario lindos problemas para resolver: desde la falta de arquitecturas que soporten implementar soluciones confiables y escalables, pasando por la configuración de redes de sensores, hasta la aplicación masiva de técnicas de machine learning que hace muchos años están consolidadas, pero que todavía no están adecuadamente implementadas en las plataformas de código abierto que usamos en nuestros proyectos.

Hoy en día, los grandes proveedores de cloud computing están implementando soluciones interesantes para integrarlas al gran volúmen de datos que están acumulando. Por ejemplo, el trabajo que hace Microsoft con Revolution Analytics y Azure Machine Learning, son extremadamente prometedores, así como las soluciones de Google y Amazon. Sin embargo, todas ellas, a pesar de lo interesante y prometedoras, todavía son plataformas que carecen de modelos que se adapten automáticamente a la casuística de una organización particular y que sean efectivas para la misma.

En la actualidad se requiere de alguien altamente especializado para que genere un modelo predictivo eficiente, que haga la diferencia para el negocio de la organización. En este sentido, aparecen startups prometedoras que venden servicios para determinadas verticales y protegen sus modelos detrás de la venta del servicio, llevando los datos del cliente a su entorno de procesamiento. No obstante, deberíamos ver una evolución en el sentido inverso, donde se puedan portar y adaptar los modelos en el entorno de ejecución del cliente, llevando el procesamiento a los datos y sirviéndose de la nube como una fuente más de información, donde las empresas del mismo rubro aportan anónimamente sus datos para evolucionar los modelos predictivos y donde se integran fuentes de datos que tiene sentido procesarlas en la nube en su propia fuente, tales como dato de redes sociales, por ejemplo.

EVOLUCIÓN DE PLATAFORMAS TECNOLÓGICAS

Estamos viendo madurar y evolucionar diversas plataformas tecnológicas. En particular, hace algunos años, tecnologías como Hadoop competían con otras plataformas, mientras los grandes de la industria desarrollaban sus propios modelos de computación distribuida, y hoy en día esa discusión ya es cosa del pasado. Así como también lo fue en un momento la existencia de diversos protocolos de redes LAN, hasta que se impuso el TCP/IP, hoy pasa algo similar con Hadoop y todo el ecosistema que corre sobre él: Storm, Spark, cascading, Scala, Hive, Hbase, etc. Vemos varias distribuciones consolidadas, la separación del almacenamiento de los data nodes, la evolución basada en yarn, todas las cuales permiten crear soluciones de órdenes de magnitud más eficientes y pensar en soluciones que operen en tiempo real.

Estas distribuciones permiten a personas con poco entrenamiento aprender la parte de manipulación de datos de Hadoop con sólo bajar una imagen de máquina virtual, trabajar con una solución que se despliega muy rápido y que permite, en las versiones de nube, almacenar los datos con el cluster apagado, y prenderlo para procesar los datos sólo cuando uno lo necesita (en algunos casos, pagando por minuto procesado).

Esta tecnología se ha convertido en algo barato y abundante, además de tener un ritmo de evolución vertiginoso. La plataforma va a seguir evolucionando más y más rápido, incluso de forma más apresurada de lo que las organizaciones pueden digerir. Tenía un profesor en la universidad que solía decir “la historia acelera”. Creo que es muy cierto.

La abundancia de capacidad de cómputo y de almacenamiento a costos muy bajos en las soluciones cloud, más el alineamiento de la comunidad académica y la industria a nivel mundial, para evolucionar sobre plataformas como Hadoop,

genera, sin dudas, espacios para que aparezcan las soluciones que logren explotar esos datos, aportando con una diferencia significativa al negocio.

Sin embargo, si vuelvo a mis tiempos de arquitecto de una organización, hay dos decisiones de arquitectura, tres fundaciones sobre las cuales construiría la estrategia de datos de la organización y que me parece son elementos previos de los cuales carecen la mayoría de las organizaciones hoy.

Primero, almacenaría todos los datos que generan los sistemas de la organización en su forma original, sin ningún tipo de transformación y en la forma más sencilla posible. Todo, no se descarta ni resume, nada.

Segundo, no permitiría que se comprara ningún sistema que no entregue, en forma de datos abiertos (datos de los cuales se puede interpretar su significado en forma integral), todos y cada uno de los datos que captura de mi organización. Y éste, aunque parezca menor, es un punto crítico hoy en Chile. Al realizar un proyecto a una minera importante y cuando pide la telemetría de algo tan simple como una pala mecánica y dice que no, sólo el proveedor puede leer los datos de la máquina, si se conecta algo a la interfaz de la computadora de la máquina se pierde la garantía. Y en último punto, no sólo se requiere de hardwares, sino también es indispensable que todo esté en la nube.

En este tiempo donde el más desinformado de los gerentes entiende que hay valor en los datos, en el concepto de Big Data, la situación de datos abiertos es inadmisible y es una decisión innegociable que tiene que imponer la arquitectura de la información; no aceptar islas de datos dentro de la organización, no aceptar proveedores que se quedan con nuestros propios datos y no nos los entreguen.

Antes hablaba de lo que se necesita para hacer data science. En un nivel más básico, para hacer machine learning se necesitan datos, patrones existentes en los datos y que estos no sean formulables en forma matemática. Bueno, estas son las tres bases para hacer machine learning, pero la más básica de todas es… ¡tener datos!

LA EXPERIENCIA EN TELEFÓNICA I+D

Nosotros trabajamos con una visión muy pragmática y partimos de problemas reales, enfocados principalmente en los desafíos referidos de minería, agricultura y urbanismo que enfrentan nuestros clientes (Market Pull). Sobre esos inconvenientes, nos preocupamos de capturar toda la información existente y de integrar sensores para capturar y poner toda la información en una plataforma de Internet of Thinks (IoT).

Esto implica que cada objeto instrumentado tendrá una existencia física en el mundo real y otra virtual, contextualizada y conectada con la instancia física, existiendo dentro de nuestra plataforma. Adicionalmente, capturamos toda la historia de cada elemento y agregamos la información de sistemas existentes y fuentes externas que sean relevantes para darle contexto al elemento.

Podemos tener la telemetría de un camión, pero es relevante para el contexto tener, por ejemplo, la condición climatológica en el lugar por donde circula el camión. Los equipos de desarrollo trabajan en los problemas de integración, visualización de los datos de los sensores, integraciones diversas y hasta el procesamiento de eventos complejos de primer nivel.

Los datos nos hablan. Por ejemplo, si aumentan las señales alternadas en un período de tiempo es indicación de que el sensor está haciendo mal contacto; un aumento repentino de consumo de energía implica una manguera de presión con pérdidas; un aumento sostenido del consumo de energía, en general indica falta de lubricación, etc.

LA APERTURA DE LOS DATOS

Otro tema muy importante que mencioné al comienzo es la disponibilidad de los datos. Respetando en primera instancia los elementos de privacidad de datos, siempre hay espacios para compartir información.

Esta arista sobre compartir los datos es muy clara y común en el caso de las ciudades: hay múltiples iniciativas de ciudades digitales y todas tienen un componente de datos abiertos. Sin embargo, a nivel empresarial no existe casi apertura y si bien esto obedece en general a políticas internas de seguridad de la información, hay datos que son valiosos por su transversalidad y porque al integrarlos, esto genera redes de información que aportan más valor que las islas (nodos) que la componen. Por ejemplo, si tengo una estación agroclimatológica, debería ser fácil para mí compartir los datos de la misma y para otros, encontrar esos datos y consumirlos, porque el poder de ajuste de predicciones que da el tener las estaciones de mis vecinos y de los vecinos de mis vecinos, es enorme.

Si somos capaces de ver el valor de la información agregada y compartida, aparecerá la necesidad de tener una plataforma transversal, e idealmente abierta, en donde agregarla y explotarla. Por ello, en Telefónica I+D montamos todas las soluciones sobre la plataforma FIWARE. Esta plataforma es creada como proyecto abierto por la comunidad económica europea, y si bien hoy la discusión de software comercial o abierto es un tema casi del pasado, sí es relevante el tema de tener la facilidad de manejar datos abiertos, es decir, que mis datos estén en un formato que los pueda consultar e interpretar tanto un humano, como una máquina, con acceso simple y basado en tecnologías estandar (acceso REST, transmisión sobre http, encapsulación en json, etc).

Para nosotros resulta esencial el poder compartir los datos abiertos que dispongamos. Si queremos tener un ecosistema de IoT activo y dinámico, es necesario que podamos darle de forma sencilla acceso a los datos a los emprendedores. El poder de entregar permiso de lectura al usuario "Juan" sobre todos los sensores del tipo "estación de clima" del grupo "Región de Coquimbo", y que con esta simple acción le demos todas las facilidades de consulta y procesamiento, es una de las capacidades "out of the box" que queremos tener en corto plazo con esta plataforma. ■