MESA REDONDA

El valor de los datos de las AAPP y su valor como herramienta para las mejoras de los servicios públicos.

Vivimos inmersos en la era de la Sociedad de la Información, en la que los datos representan el verdadero valor. Primero fue el patrón oro. Luego empezamos a hablar del petróleo como el "oro negro" que dominaba la economía. Hoy el nuevo "petróleo" son los datos.

D. ALFONSO CASTRO
Director del Departamento de Informática Tributaria

Prueba de ello es constatar que cinco de las seis empresas de mayor capitalización mundial son Tecnológicas (1- Apple, 2-Google, 3-Microsoft, 4 Exxon, 5-Amazon, 6-Facebook).

Esto es especialmente aplicable a las AAPP ya que es esencial para nuestra función el adecuado manejo de la información. Las AAPP están en un acelerado proceso de digitalización, y si en el pasado pudimos producir algún producto tangible (recordemos el BOE papel, o el disquete con del programa Padre que se vendía en los estancos) estos y otros se han desmaterializado al pasar a formatos electrónicos.

En el caso concreto de la AEAT, los datos constituyen nuestra verdadera materia prima. Básicamente, la labor de la AEAT es el procesamiento de información (captación de declaraciones, cruces, generación de actos administrativos...), por tanto, los datos son la **savia** que permite funcionar a la organización. Los datos son la materia prima que es necesario recopilar y mantener.

Podemos decir que la AEAT tiene dos grandes objetivos: Asistencia (reducción de los costes indirectos derivados de las obligaciones formales) y Control (lucha contra el fraude). Este uso de los datos, nos apoya de una manera evidente en la función de Control, pero, aunque parezca menos evidente, también en la Asistencia, basta pensar en nuestro proyecto estrella, la campaña de Renta, donde a partir de la información de terceros la AEAT proporciona un borrador de declaración (este año disponible para todos los contribuyentes). Esto tiene un do-

ble valor: a los contribuyentes se les simplifica su labor, pero por otro lado estimula el correcto cumplimiento de forma voluntaria, lo que mejora la recaudación y por tanto nuestra eficacia.

Es importante centrarse en los datos, en cómo gestionarlos como activo y convertirlos en un valor tangible para el negocio.

Los datos son reutilizables, **nunca se agotan**, y tienen múltiples aplicaciones y lecturas. Los datos que un departamento utiliza con un fin determinado, en otro punto de la organización tienen otra utilidad. La utilidad puede ser instantánea en el momento de la captura del dato o derivada en el tiempo como comportamiento histórico.

Existe una necesidad real, y cada vez más acuciante en todas las empresas, de organizar la información. Multitud de datos, informes, registros, canales, medios y procesos en los que se contiene y fluye la información, se multiplican de forma imparable y crecen sin control. Sin una visión integral de los datos, se crean silos de información por departamentos o áreas funcionales.

Durante años hemos invertido en las aplicaciones operacionales, pero apenas en la organización de los datos, cuando son los datos los que finalmente aportan el conocimiento que crea valor y negocio en las organizaciones.

Los datos son un activo corporativo, necesitamos profesionalizar el tratamiento de la información y los datos desde la óptica de negocio. Debemos, por tanto, abordar la **gobernanza de la información**, en especial en estas cuatro facetas: **completitud**, **seguridad**, **calidad y claridad semántica.** Estos factores son esenciales en cualquier organización que pretenda sacar provecho de sus datos, y esto aplica independientemente de la tecnología con la que se aborde esta tarea.

Proyectos de procesamiento de grandes volúmenes de información para aportar valor a vuestros servicios en los trabaja la AEAT.

Nos referimos al uso de técnicas modernas de tratamiento masivo de una información caracterizada por las denominadas "tres uves": **volumen, velocidad y variedad.**

Estas técnicas permiten utilizar hardware convencional y software especializado para distribuir grandes volúmenes de datos entre diferentes unidades de proceso de forma que éstas pueden gestionar en paralelo la parte que les corresponde de los datos para luego, a partir de estos resultados parciales, obtener un resultado total. Esta estrategia resulta escalable, es decir, permite afrontar el cada vez más creciente volumen de datos disponibles a tratar mediante un crecimiento "ilimitado" de almacenamiento y unidades de proceso a un precio moderado.

Sin embargo, el uso de la expresión Big Data se ha extendido y abarca generalmente otras facetas del tratamiento de la información como la analítica avanzada de datos o la extracción y análisis de información disponible en fuentes abiertas (Internet).

La ambigüedad de la expresión es tal que incluso lo que para una organización es un gran reto tecnológico y puede considerarse un claro ejemplo de Big Data, para otra puede estar superado hace años. De hecho, realmente no son tan numerosas las organizaciones que tienen a su disposición cantidades de información.

Con el objetivo de explorar tecnologías de almacenamiento y procesamiento distribuidos pusimos en marcha un proyecto de Big Data (en sentido estricto) en la AEAT.

Tras confirmar la gran capacidad de escalado y la posibilidad de obtener tiempos de respuesta ágiles en el tratamiento de estos volúmenes de datos en comparación con la tecnología usada hasta ahora, seleccionamos las tecnologías más adecuadas para nuestra necesidad dentro del conjunto de herramientas del llamado ecosistema *Hadoop*, referente habitual para el tratamiento de volúmenes masivos de información.

A modo de ejemplo tenemos cargado en *Hadoop*, para su análisis usando Impala, **40 mil millones de registros (4*10^10)** correspondientes a los HITs de nuestra SEDE de los últimos 4 años.

Esto nos ha permitido confirmar la fortaleza del ecosistema *Hadoop* para escalar casi linealmente con respecto al número de nodos dedicados al cluster (tanto en lo que se refiere al tamaño de las fuentes, como al número de peticiones atendidas por segundo).

Otros ejemplos de uso del ecosistema *Hadoop* en la AEAT son:

- *Spark-GraphX (Algoritmo Pregel)* para determinación de la riqueza societarias. Se analiza la red de relaciones (personas que tienen acciones de empresas que a su vez tienen acciones de otras empresas), recorriendo todos los caminos para determinar la riqueza societaria imputable a cada persona.

- *Lucene* para búsqueda en textos. Cerca de 10 millones de documentos indexados, muchos de los documentos proceden de escaneado de papeles presentados por el contribuyente en cuyo caso previamente se realiza OCR del mismo (45 millones páginas realizado OCR). Hemos incluido *Lucene* en esta relación, ya que a pesar de no ser parte del ecosistema *Hadoop*, lo cierto es que hemos verificado que escala de forma adecuada al añadir nodos al cluster.

Zújar, el datawarehouse de la AEAT.

Si bien el uso de herramientas del ecosistema *Hadoop* ha demostrado su versatilidad, lo cierto es que en la AEAT se lleva muchos años invirtiendo en crear un DataWarehouse corporativo, denominado *Zújar*, que con más de mil Fuentes de Datos representa el verdadero corazón de la analítica de la AEAT. A través de esta herramienta toda la organización tiene a su disposición de forma ágil y sencilla la información necesaria.

Con *Zújar*, y de forma interactiva, los usuarios pueden buscar y visualizar los datos, ordenarlos, filtrarlos, agruparlos, sumarlos, compararlos... Sobre este sistema se hacen los contrastes de información que permiten detectar inconsistencias entre diferentes fuentes de datos (por ejemplo, el IVA que una empresa declara haber soportado frente al IVA que otras empresas manifiestan haberle repercutido en sus declaraciones).

Desde el punto de vista tecnológico, la herramienta es un desarrollo propio que se ha apoyado históricamente en un sistema de bases de datos orientado a columnas. Este sistema ofrece tiempos de respuesta ágiles para el tratamiento libre de informaciones muy voluminosas, pero los tiempos se deterioran cuando la fuente supera los mil millones de registros.

Para fuentes mayores usamos, como ya hemos comentado, *Hadoop-Impala* pero con la misma interfaz gráfica del *Zújar* debido a su potencia y sencillez de uso, así como al conocimiento que ya hay de la misma dentro de la organización. Esto ha exigido un esfuerzo de integración entre ambos sistemas, pero a cambio abre el uso inmediato de la plataforma Big Data a toda la AEAT. Hay que resaltar que el acceso desde *Zújar* a la mayor parte de las fuentes de datos mantiene la tecnología original por tener otras fortalezas cuando los volúmenes de información no son tan masivos (por debajo de mil millones de registros).

Para entender en qué medida la herramienta *Zújar* está integrada en la cultura de la organización basta señalar que cuenta con más de 10.000 usuarios diferentes (alrededor del 40% de los trabajadores de la AEAT y otros usuarios externos autorizados), y se usa extensivamente no sólo para la selección y análisis de contribuyentes en la lucha contra el fraude, sino también para realizar análisis estadísticos y seguimiento de la actividad de la organización a todos los niveles.

El éxito de esta herramienta se debe en parte a algunas características de un desarrollo propio difícilmente reproducibles por productos de mercado, como su facilidad de uso para el análisis de información tributaria o su integración con las aplicaciones gestoras, de forma que es posible realizar selecciones en las herramientas de análisis que determinan el colectivo sobre el que se aplicarán actuaciones de tramitación individuales o masivas en las aplicaciones de gestión.

La seguridad de *Zújar*, integrada con la de los sistemas operacionales de la AEAT, tiene una granularidad que permite asegurar que cada usuario sólo pueda acceder a aquella información que realmente es relevante para su trabajo, ya sea desde un punto de vista funcional, territorial, o incluso por las características de los datos a los que quiere tener acceso. Esto permite mantener el control de lo que se puede o no hacer independientemente del número de usuarios y la variedad de los datos.

Sin embargo, dada la riqueza semántica de la información a disposición de cualquier miembro de la organización, sería muy difícil para ellos la comprensión del modelo de datos subyacente si no se hubiera realizado una ardua tarea de modelización de la información que permite a cada usuario entender el dato que está viendo y buscar qué informaciones son de interés en términos de negocio y no técnicos. Toda esta información, volcada en un diccionario de metadatos se convierte en el corazón del sistema.

"El éxito de esta herramienta *[Zújar]* se debe en parte a algunas características de un desarrollo propio difícilmente reproducibles por productos de mercado, como su facilidad de uso para el análisis de información tributaria o su integración con las aplicaciones gestoras."

Además, aunque los usuarios tuvieran la información disponible, de poco serviría si no pudieran confiar en ella por no estar seguros de que tiene la calidad adecuada. Los procesos de preparación de la información para incorporación a la herramienta de análisis al igual que la modelización de los datos se supervisan por los responsables de las aplicaciones operacionales correspondientes y por los propios responsables de negocio de tal forma que la organización no tiene dudas de la información analizada.

El sistema se complementa con un generador de informes normalizados (en nuestro caso, un desarrollo propio bautizado como *Genio*), que permiten predefinir análisis complejos. Esto es especialmente útil cuando es claro el diagnóstico que se busca (según el paradigma del análisis de sangre, que concentra en un documento un conjunto de pruebas realizadas resaltando aquellas que ofrecen resultados indicativos de anomalías) y es de gran utilidad y efectividad, por ejemplo, en la selección de contribuyentes para comprobación, donde aporta además el valor adicional de la objetividad en el tratamiento de los contribuyentes.

Los párrafos anteriores resumen como la AEAT aborda la **gobernanza de la información (completitud, seguridad, calidad y claridad semántica)**, algo esencial en cualquier organización que pretenda sacar provecho de sus datos.

Principales oportunidades encontradas para mejorar los servicios mediante estas técnicas.

Las oportunidades son innumerables:

- **BigData:** Mejor aprovechamiento de la información y posibilidad de abordar análisis que hasta el momento eran imposibles.

- **Analítica Avanzada:** Descubrimiento de patrones y creación de modelos.

- **Búsqueda en fuentes abiertas:** Enriquecimiento del Sistema con información pública y poder correlacionarla.

Todo lo anterior aumenta la eficacia de la Administración al permitirnos hacer más con menos dedicando los RRHH a temas que aporten mayor valor a la organización, aumentando la productividad y, en el caso de la AEAT, incrementando la cifra alcanzada en la lucha contra el fraude.

Principales problemas encontrados:

- Dificultad para encontrar, junto a las unidades de negocio, los casos de uso consiguiendo la necesaria implicación del negocio.

- Falta de especialistas en el mercado.

- Ecosistema de herramientas muy amplio, lo que aumenta la dificultad para seleccionar la mejor herramienta para cada caso.

La figura del Chief Data Officer (CDO) en la Administración. ¿Es necesaria? ¿Debe ser TIC?

En nuestro caso hemos optado por otro modelo. Cada responsable de una aplicación es responsable por un lado del operacional (la tramitación) y por otro de definir el DataWarehouse de las fuentes de datos de dicha aplicación.

Adicionalmente existe un área encargada de la carga del Datawarehouse, pero no podemos decir que el responsable de esa área corresponda con la figura del CDO. En mi opinión, de existir el CDO, debería ser un TIC que dependiese del CIO. ✱