Algunos conceptos de Arquitectura de Datos asociados con Big Data

Fernando López-Torrijos

Noviembre de 2021

Introducción

DAMA International (https://www.dama.org/cpages/home) es una organización internacional, con capítulos por países, que se dedica a hacer avanzar los conceptos y las prácticas de la gestión de la información y de los datos y a apoyar a sus miembros para que aborden sus respectivas necesidades de gestión de la información y de los datos con las mejores prácticas conocidas.

Es así que publican el DMBOK (Data Management Body of Knowledge)

El capítulo 14 del libro trata de lo relacionado con Big Data y Data Science. Este capítulo incorpora todo lo planteado en los capítulos anteriores respecto a la Arquitectura y Gestión de Información y Datos (Administración, Gobierno, Arquitectura, Modelado y Diseño, Almacenamiento y Operación, Seguridad, Integración e Interoperabilidad, Administración documental, Datos Maestros y de Referencia, Data Warehousing y Business Inteligence, Gestión de metadatos y Gestión de la Calidad de los Datos), y puntualiza temas importantes o diferenciadores. Se realiza a continuación un resumen.

La estrategia de Big Data.

Las preguntas que debe resolver el diseño de una estrategia de negocio que incorpore Big Data son:

¿Qué problemas intenta resolver la organización? ¿Para qué necesita la analítica del Big Data?

Aunque una ventaja de la Ciencia de Datos es que puede proporcionar una nueva perspectiva a una organización, ésta necesita tener un punto de partida. Una organización puede determinar que los datos se van a utilizar para entender el negocio o el entorno empresarial; para probar ideas sobre el valor de nuevos productos; para explorar algo desconocido; o para inventar una nueva forma de hacer negocios. Es importante establecer un proceso de entrada para evaluar estas iniciativas en varias fases durante la implementación. El valor y la viabilidad de las iniciativas debe evaluarse en varios momentos. Por ende, deben tenerse métricas asociadas que permitan realizar dicha evaluación.

¿Qué fuentes de datos utilizar o adquirir?

Las fuentes internas pueden ser fáciles de utilizar, pero también pueden tener un alcance limitado. Las fuentes externas pueden ser útiles, pero están fuera del control operativo ya que son gestionadas por terceros, o no controladas por nadie, como en el caso de las redes sociales. Muchos proveedores compiten en este espacio y a menudo existen múltiples fuentes para los elementos o conjuntos de datos deseados. Adquirir datos que se integren con elementos de *ingesta* existentes puede reducir los costes globales de inversión.

¿Cuál es la oportunidad y el alcance de los datos a suministrar a la analítica?

Pueden proporcionarse en tiempo real, por medio de instantáneas en un momento dado, o integrados y resumidos.

Los datos de baja latencia son ideales, pero hay una gran diferencia entre los algoritmos computacionales dirigidos a los datos en reposo frente a aquellos dirigidos a analizar *flujos de datos*. No hay que obviar el esfuerzo adicional que implica integrarse con *flujos de datos*.

¿Cuál será el impacto y la relación con otras estructuras de datos presentes en la organización?

Es posible que haya que cambiar la estructura o el contenido de estructuras de datos actuales para hacerlas aptas para su integración con conjuntos de Big Data.

¿Habrá influencia en los modelos de datos existentes?

Incluya la perspectiva de que habrá una ampliación del conocimiento sobre clientes, productos y enfoques de marketing.

La estrategia elegida, es decir, las respuestas a estas preguntas, afectará el alcance y el calendario de la hoja de ruta de las capacidades de Big Data de una organización.

Alineación de la estrategia.

Cualquier programa de Big Data debe estar estratégicamente alineado con los objetivos de la organización. El establecimiento de una estrategia de Big Data genera actividades relacionadas con la comunidad de usuarios, la seguridad de los datos, la gestión de metadatos, incluido el *linaje*[1], y la gestión de la calidad de los datos.

La estrategia debe documentar los objetivos, el enfoque y los principios de gobernanza.

La capacidad de aprovechar la Big Data requiere de la creación de habilidades y capacidades organizativas. Utilice la *gestión de capacidades* de la organización para alinear las iniciativas de negocio y de TI y proyectar una hoja de ruta.

Los entregables de la estrategia deben dar cuenta de la gestión de:

- El ciclo de vida de los datos.
- Los metadatos.
- La calidad de los datos.
- La adquisición de datos.
- El acceso y seguridad de los datos.
- El gobierno de los datos.
- La privacidad de los datos.
- El aprendizaje y la adopción.
- Las operaciones.

Como en cualquier proyecto de desarrollo, la implementación de una iniciativa de Big Data debe alinearse con las necesidades reales del negocio. Evaluar la preparación de la organización en relación con los factores críticos de éxito:

[1] Es el proceso de comprensión, registro y visualización de los datos a medida fluyen desde las fuentes de datos hasta su consumo explicando las transformaciones que sufre a lo largo del camino, qué cambia y por qué. A veces lo denominan cadena de datos.

- Relevancia para el negocio: ¿En qué medida las iniciativas de Big Data y sus casos de uso correspondientes se alinean con el negocio de la empresa?
- Preparación del negocio: ¿Cuál es la brecha media de conocimientos o habilidades dentro de la organización? y ¿Puede ser superada en un solo incremento?
- Viabilidad económica: ¿La solución propuesta ha considerado de forma conservadora los beneficios tangibles e intangibles? ¿La evaluación de los costes de propiedad ha tenido en cuenta la opción de comprar o alquilar elementos frente a construir desde cero?
- Prototipo: ¿Puede la solución propuesta ser objeto de un prototipo durante un plazo limitado para demostrar el valor propuesto? Las implantaciones de gran envergadura pueden causar grandes impactos monetarios y un espacio para pruebas puede mitigar estos riesgos.

La táctica del Big Data

En la implementación de la estrategia, la táctica, se aplican a la gestión de Big Data muchos de los principios generales de la gestión de Data Warehouses:

1. Garantice que las fuentes de datos sean fiables.
2. Tenga suficientes metadatos para permitir el uso de los datos.
3. Gestione la calidad de los datos.
4. Averigüe cómo integrar los datos de diferentes fuentes.
5. Garantice que los datos estén seguros y protegidos.

Las diferencias en la implementación de un entorno de Big Data están relacionadas con un conjunto de incógnitas: ¿Cómo se utilizarán los datos?, ¿Qué datos serán valiosos?, ¿Cuánto tiempo hay que conservarlos?

El stremaing puede llevar a la gente a pensar que no tienen tiempo para implementar controles. Es una suposición peligrosa. Con conjuntos de datos más grandes, la gestión de la ingestión y el inventario de datos en un *Data Lake* es fundamental para evitar que se convierta en un *Data Swamp*.

Las fuentes de datos.

Al igual que con cualquier proyecto de desarrollo, la elección de las fuentes de datos para el trabajo de Ciencia de Datos debe ser impulsada por los problemas que la organización está tratando de resolver. La diferencia con el desarrollo de Big Data para Data Science es que la gama de fuentes de datos es más amplia. No está limitado por el formato y puede incluir datos tanto externos como internos a una organización. La capacidad de incorporar estos datos a una solución también conlleva riesgos. La calidad y la fiabilidad de los datos deben evaluarse y debe establecerse un plan de uso a lo largo del tiempo. Los entornos de Big Data permiten ingerir rápidamente muchos datos, pero para utilizarlos y gestionarlos a lo largo del tiempo sigue siendo necesario conocer los datos básicos:

- Su origen.
- Su formato.
- Qué representan los elementos de los datos.
- Cómo se conectan con otros datos.
- Con qué frecuencia se actualizará.

Elección de las fuentes de datos

Hay que revisar las fuentes de datos disponibles y los procesos que los crean y gestionar un *plan de nuevas fuentes*.

Tenga en cuenta:

- Datos propios (por ejemplo los puntos de venta).
- Granularidad: Lo ideal es obtener los datos en su forma más granular. Luego se podrán agregar si se requiere.
- Coherencia: Seleccione conjuntos de datos cuyos datos sean coherentes entre sí.
- Fiabilidad: Elija fuentes de datos que sean fiables, autorizadas y creíbles a lo largo del tiempo.
- Inspeccionar/perfilar[2] las nuevas fuentes: Pruebe los cambios antes de añadir nuevos conjuntos de datos. Pueden producirse cambios inesperados o cambios significativos en los resultados de las visualizaciones o los cuadros con la inclusión de nuevas fuentes de datos.

Es necesario evaluar el valor y fiabilidad de cada fuente.

Los riesgos asociados a las fuentes de datos incluyen los temas legales y éticos, como la *privacidad*.

Una vez identificadas las fuentes, a veces hay que gestionar la compra, e ingerirlas (cargarlas) en el entorno de Big Data. Durante este proceso se capturan los metadatos críticos sobre la fuente, como su origen, tamaño, tipo y conocimientos adicionales sobre el contenido. Muchos motores de ingesta perfilan los datos a medida que son ingeridos, proporcionando a los analistas por lo menos metadatos parciales. Una vez que los datos están en un *Data lake*, se puede evaluar su idoneidad para múltiples esfuerzos de análisis. Dado que la construcción de modelos de Ciencia de Datos es un proceso iterativo, también lo es la ingesta de datos.

Identifique periódicamente las carencias a partir del inventario de activos de datos e incorpore esas fuentes.

Antes de integrar los datos, evalúe su calidad. La evaluación puede consistir en una simple consulta para averiguar cuántos campos contienen valores nulos, o en una tan compleja como ejecutar un conjunto de módulos de calidad de datos para perfilar, clasificar e identificar las relaciones entre los elementos de los datos. Esta evaluación permite saber si los datos proporcionan una muestra válida a partir de la cual trabajar y, en caso afirmativo, cómo se pueden almacenar (dispersos en unidades de procesamiento lógico, federados, distribuidos por clave, etc.). El proceso de evaluación proporciona una valiosa visión de cómo se pueden integrar los datos con otros conjuntos de datos, con los datos maestros y/o con los datos históricos. También será útil para la etapa de entrenamiento de modelos y para las actividades de validación de éstos.

La *ingesta.*

La capacidad de ingerir e integrar rápidamente datos de diversas fuentes a gran escala permite unir conjuntos de datos que de otro modo estarían aislados. Por ejemplo, centralizar la información de todos los puntos de venta a nivel nacional (o internacional). Se podrá realizar un análisis a través de un resumen, un agregado o un modelo y divulgarlo. Pero también se podría analizar a un nivel geográfico local o regional muy específico y divulgarlo con los datos anonimizados si fuere necesario.

Los criterios utilizados para seleccionar o filtrar los datos suponen un riesgo. Hay que evitar sesgos y discriminaciones. El filtrado puede tener un impacto en el análisis. El discernimiento respecto a las consecuencias derivadas de la eliminación de valores atípicos es necesario, así como en las ocasiones en que se restringen los conjuntos de datos a un dominio limitado o se eliminan los elementos dispersos.

[2] Análisis de los datos evaluando su calidad e identificando la conformidad con los requerimientos de calidad que se le exigen.

La mayor diferencia entre el procesamiento en un Data Warehousing/Business Inteligence y Big Data es que en un almacén de datos tradicional los datos se integran a medida que se introducen en el almacén: ETL (extraer, transformar, cargar), mientras que en un entorno de Big Data, los datos se ingieren y cargan antes de ser integrados: ELT (extraer, cargar, transformar).

En Big Data los datos podrían no estar integrados en absoluto, en el sentido tradicional. En lugar de integrarse como preparación para su uso, a menudo se integran a través de usos ad hoc (por ejemplo, el proceso de construcción de modelos predictivos impulsa la integración de conjuntos de datos concretos).

Esta diferencia entre ETL y ELT tiene importantes implicaciones en la gestión de los datos. Por ejemplo, el proceso de integración no se basa necesariamente en la producción de un *modelo de datos empresarial*. El riesgo es que se puede perder mucho conocimiento sobre los datos si los procesos de ingestión y uso se ejecutan de forma ad hoc. Es necesario recopilar y gestionar los metadatos relacionados con estos procesos, si se quiere que se entiendan y se aprovechen a lo largo del tiempo.

La ingestión no siempre requiere la propiedad de la organización. Considere la posibilidad de alquilar una plataforma de Big Data por períodos finitos para explorar los datos de interés. La exploración puede determinar rápidamente qué áreas muestran un valor potencial. Haga esto antes de la ingesta en el *Data Lake* de la organización, el *Data Warehouse* u otra área de almacenamiento; una vez que se ha incorporado, puede ser difícil de eliminar.

El modelado.

El modelado de Big Data es un reto técnico pero crítico si una organización quiere describir y gobernar sus datos. Los principios tradicionales de la Arquitectura de Datos Empresarial se aplican; los datos deben ser integrados, especificados y gestionarse.

El principal impulsor para el modelamiento físico de un almacén de datos es permitir el rendimiento de las consultas. Que sea algo muy técnico no es una excusa para dejar el proceso de modelado en manos de un desarrollador. El valor de modelar los datos es que permite a las personas entender el contenido de los datos y por ello la participación de varias áreas es enriquecedor para la organización.

Desarrolle de uno en uno el modelo de cada área temática. Hágalo al menos de forma resumida para que pueda relacionarse la nueva Big Data con las entidades contextuales adecuadas y colocarse en la *hoja de ruta* general, al igual que cualquier otro tipo de datos. El reto consiste en obtener una imagen comprensible y útil de estos grandes volúmenes de datos, y a un coste justificable.

El modelado debe entender cómo se enlazan los datos entre conjuntos de datos. En el caso de los datos de distinta granularidad, hay que evitar combinar conjuntos atómicos y agregados.

Los roles

Al igual que con el Data Warehouse/Business Inteligence, una implementación de Big Data reunirá una serie de roles cruzados clave, incluyendo:

- Arquitecto de la plataforma de Big Data: Hardware, sistemas operativos, sistemas de archivos y servicios.
- Arquitecto de ingestión: Análisis de datos, sistemas de registro, modelado de datos y mapeo de datos. Proporciona o apoya la asignación de fuentes al clúster Hadoop para su consulta y análisis.
- Especialista en metadatos: Interfaces de metadatos, arquitectura de metadatos y contenidos.
- Jefe de diseño analítico: Diseño analítico para el usuario final, implementación de las mejores prácticas de herramientas relacionadas, y facilitador del conjunto de resultados al usuario final.