MESA REDONDA

El valor del dato.

Fue un placer participar en las jornadas AsticNET celebradas el pasado mes de Junio y comprobar como los términos como Hadoop, Analítica predictiva, NLP, etc... forman parte ya del día a día de las organizaciones TIC en la Administración del Estado.

D. IGNACIO ARRIETA
Director de la Preventa
DellEMC Iberia

Siendo esto así, se percibe que en las organizaciones TIC hay un estado de diferente madurez en lo referente a la soluciones de analítica/BigData: desde la aplicación de técnicas de proceso paralelo a pequeños casos de uso, hasta grandes plataformas que permiten la autoprovisión de datasets y herramientas concretas a diferentes grupos. Independientemente del estado de madurez de la organización, es de crucial importancia el tomar la dirección correcta en lo que se refiere a la estrategia de arquitectura para que las organizaciones puedan extraer el máximo del dato, tanto ahora como en el futuro.

En general, las arquitecturas tradicionales para sistemas de analítica/BigData apostaban por la utilización de sistemas cerrados, en los que el cada nodo del sistema aportaba capacidades de cómputo y almacenamiento. Estas arquitecturas asumían que el hecho de existir cierta localidad de los datos, aumentaba el rendimiento global del sistema. En general, este principio ha sido refutado por varios investigadores[1] *(figura 1)*.

Desde DellEMC creemos que la mejor estrategia que se puede adoptar en cuanto a la arquitectura para sistemas de analítica/BigData es la de creación de un Data Lake[2].

1 http://research.microsoft.com/en-us/um/people/ga/talks/disk-irrelevant.pdf

2 https://en.wikipedia.org/wiki/Data_lake

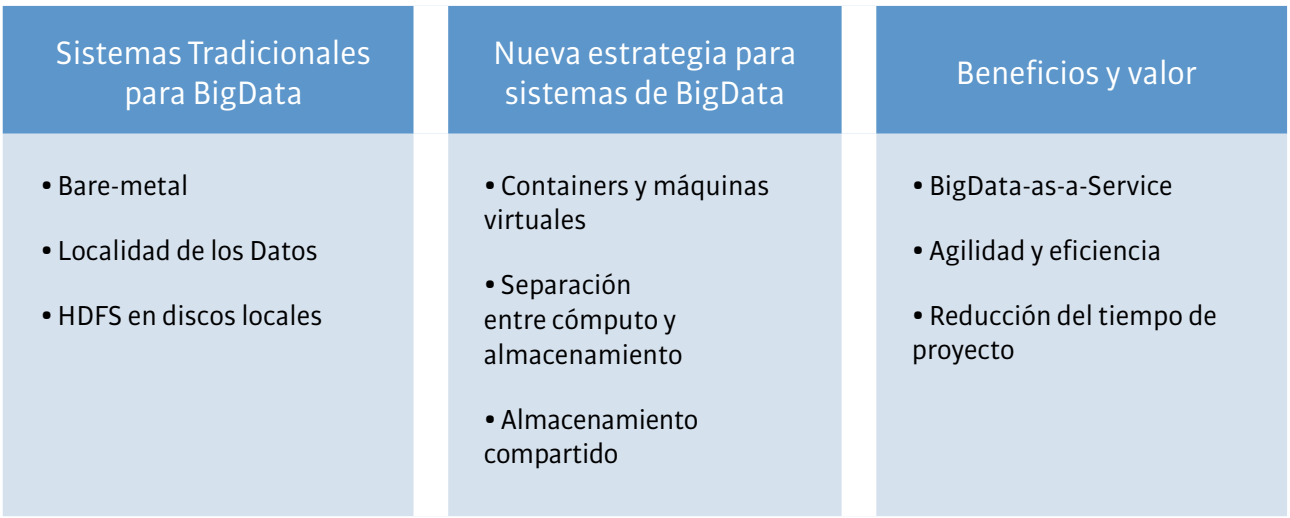

Figura 1.

Entendemos como DataLake un repositorio de datos, que habla de forma nativa varios protocolos, entre ellos el protocolo HDFS[3], que es el usado por Hadoop para almacenar datos. Este protocolo se ha convertido en un estándar de facto en lo referente a soluciones de analítica/Bigdata. El hecho de poner en juego el concepto de DataLake, permite desacoplar la capa de cómputo de la capa de almacenamiento, rompiendo el paradigma de modelo cerrado presente en las primeras versiones de arquitectura de Hadoop, en las que el almacenamiento se consumía de forma local.

Esta arquitectura en dos capas permite una mayor flexibilidad, en cuanto a que diferentes proyectos de analítica tienen diferentes requerimientos de recursos, ya sea cómputo o ya sea almacenamiento. Imaginemos, por ejemplo, el análisis en tiempo real de un flujo de logs de nuestro sistema de seguridad perimetral, donde el volumen de datos es limitado pero necesitaremos mucha potencia de cálculo, versus el análisis en modo batch de grandes datasets donde el almacenamiento será mucho más masivo.

Asimismo esta arquitectura aporta versatilidad, ya que al tratarse de un repositorio que usa el estándar HDFS podremos explotar los mismos datos desde varios sistemas y/o distribuciones, todo ello sin tener que realizar transformaciones de los datos (ETLs), ahorrando tiempo y recursos de cómputo.

Por último, otra de las consecuencias de la existencia de un DataLake es que por defecto favorece la unicidad de los datos, evitando el hecho de tener datasets duplicados.

En una arquitectura como la descrita, en la que existe una capa de cómputo segregada, tiene sentido la virtualización de los nodos analíticos. Hadoop es un sistema de proceso paralelo que con los mismos recursos disponibles, rinde mejor con un número superior de nodos[4]. Además, si desde la perspectiva de la técnica de sistemas tiene sentido el hecho de utilizar las posibilidades que ofrece la virtualización, ¿por qué no aprovecharlas también para los proyectos de analítica? *(figura 2)*

Una vez creada esta arquitectura de dos capas, con un DataLake como capa de almacenamiento y una capa de cómputo virtualizada, nos podemos plantear el hecho de ir más allá y crear una plataforma de BigData-as-a-Service. En esta plataforma podríamos automatizar y orquestar el despliegue de diferentes datasets, diferentes herramientas, ya sean de proceso y/o visualización, siempre teniendo en cuenta quien puede o no acceder a ciertos datos, estableciendo así la gobernanza efectiva de los datos. *(figura 3)*

Asimismo, esta plataforma debería de ser capaz de aprovechar de una manera ágil las conclusiones obtenidas del análisis de los datos, ya sea en forma de nuevas aplicaciones y servicios, o en forma de mejora de los servicios existentes. Poder llegar a cerrar el ciclo virtuoso en el que se crean nuevos servicios con los datos

3 https://en.wikipedia.org/wiki/Apache_Hadoop#HDFS

4 http://www.vmware.com/files/pdf/techpaper/Virtualized-Hadoop-Performance-with-VMware-vSphere6.pdf

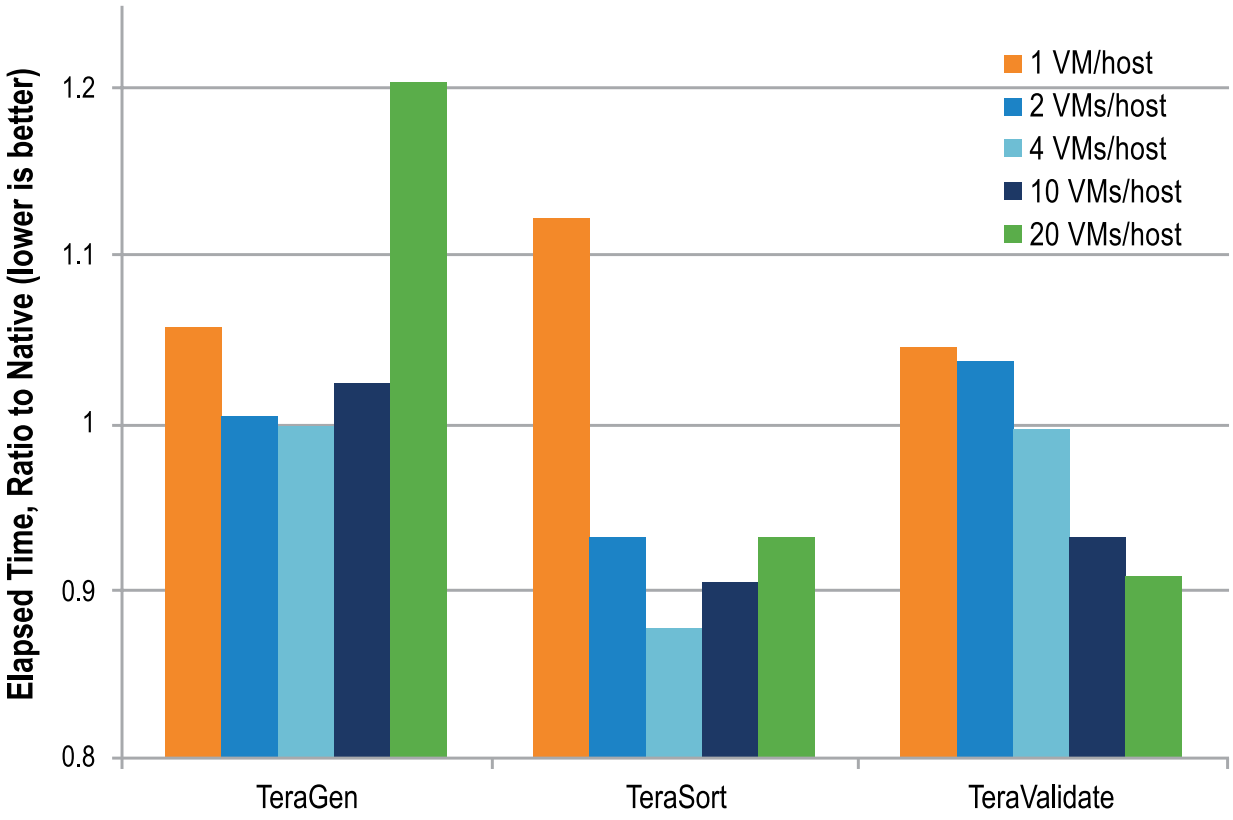

Figura 2. **Ratio of elapsed times on virtualized platforms to the native platform. Number of VMs is per host. Lower is better.**

analizados, que a su vez generan más datos que permiten mejorar estos servicios debiera ser el objetivo a largo plazo al que toda organización TIC debiera aspirar.

Si reflexionamos, la Transformación Digital de la Administración del Estado va precisamente de eso, de utilizar la tecnología disponible a día de hoy, en este caso la analítica, para mejorar la prestación de servicios a la ciudadanía, ya sea por la mejora de los servicios actuales o por la creación de nuevos. Por encima de la tecnología, esta transformación requiere de un capital humano muy concreto, tal como el personal perteneciente al Cuerpo Superior de Sistemas y Tecnologías de la Información de las AAPP, capaces de conjugar por un lado el conocimiento del "negocio", con las posibilidades que ofrece la tecnología por otro. Me gusta pensar en el concepto de traductores tecnológicos, personal empotrado en las diferentes unidades de negocio de las organizaciones, que hacen de enlace entre el negocio y las organizaciones TIC de cada organismo.

Sinceramente creo que este talento existe en vuestro Cuerpo, y tanto mi empresa, DellEMC, como yo mismo estaremos encantados de colaborar con vosotros en ese difuso camino que es la Transformación Digital. ✱

Figura 3.