BigData: ¿finalidad o medio?

por Joan Massegú
Business Analytics Solution Architect, avanttic

El “Big Data” está en un estado tan incipiente aún y hay tal cantidad de herramientas emergiendo alrededor de Hadoop, que a menudo, esa capa inicial que es la plataforma, acapara toda nuestra atención.

En este artículo vamos a hacer énfasis en las herramientas de “**Advanced Analytics**”, situadas en la cúspide de lo que llamaremos **Pirámide de Valor del Big Data**, que representa el conjunto de necesidades y/o procesos que las organizaciones van a poder necesitar en el camino hacia la extracción del máximo valor de sus datos, y cómo trasladarlo al negocio.

Big Data: ¿finalidad o medio?

Quién más quién menos, a estas alturas todos hemos oído hablar de las ya famosas tres V’s (**V**elocidad, **V**olumen y **V**ariedad), que en el mundo conectado del siglo XXI son responsables de la generación de tal cantidad de información, que los sistemas de almacenamiento y herramientas de proceso tradicionales no son capaces de tratar.

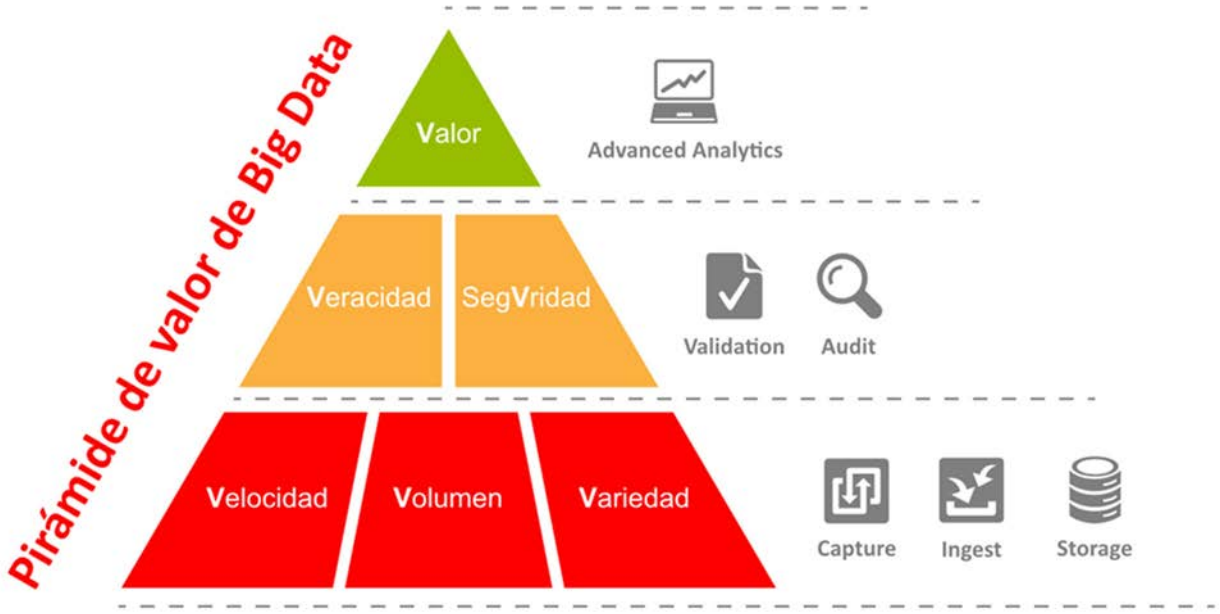
Y somos también conscientes de que ese volumen ingente de datos, contiene información que puede, de varias maneras, generar **V**alor para nuestras organizaciones… siempre y cuando seamos capaces de extraerla, asegurar una mínima **V**eracidad (y/o calidad de los datos) y custodiarla bajo las medidas de “seg**V**ridad” que su naturaleza pueda requerir.

La Pirámide de Valor del Big Data

Nuevas arquitecturas de sistemas distribuidos, basados principalmente en una plataforma (Hadoop), han sido desarrolladas para almacenar todos esos datos, y de la mano de nuevas herramientas, lenguajes y técnicas de procesamiento, permitir a las organizaciones combinar las “V” anteriormente descritas, a lo largo de diferentes fases y de diversas maneras, retroalimentándose incluso en ocasiones.

A medida que vayan avanzando en tareas de la limpieza, preparación, interconexión, etc. de los datos que se recogieron en la base, las organizaciones irán construyendo lo que llamaremos **Pirámide de Valor del Big Data.** De esta manera, irán escalando niveles que realizarán aportaciones diferentes, hasta alcanzar la cúspide, donde podrán, bien sea en forma de descubrimiento de nuevos hechos o relaciones, o la formulación de recomendaciones y/o predicciones, extraer el Valor real del Big Data.

A menudo, al hablar de Big Data, nos centramos excesivamente en el primer nivel de la pirámide, el más relacionado con la plataforma: la adquisición y el almacenamiento de la información. En el artículo de hoy vamos a concentrarnos en el nivel superior de la pirámide, donde las herramientas de **Advanced Analytics** permiten trasladar la información obtenida en la base de la pirámide al negocio, con una calidad en contenido y formato que posibilite a las organizaciones extraer el valor del Big Data, marcando así una diferencia competitiva respecto a su competencia.



Tipos de Análisis

Vamos ver a continuación los diferentes tipos de análisis posibles y las herramientas de **Advanced Analytics** que Oracle nos ofrece para cada uno de ellos.

El **análisis descriptivo** es aquél que permite explicar las cosas que ya han pasado y está ligado al análisis diagnóstico, que explica por qué han pasado. Ambos tipos de análisis caben bajo el paraguas de lo que hasta ahora hemos conocido como Business Intelligence (aunque éste estaba restringido a información estructurada almacenada en un DWH, o incluso en forma de reporting contra el OLTP).

El **análisis predictivo** nos ayuda a avanzarnos a lo que va a ocurrir para poder tomar decisiones competitivas de manera informada (en base a información recolectada por diferentes métodos, sea o no estructurada), mientras que el **análisis prescriptivo** es el que nos recomienda lo que deberíamos hacer (Decision Support Systems – DSS), llevándonos incluso a escenarios de organizaciones "data-driven", dónde podrían tomarse determinadas decisiones de manera automática.

Herramientas Oracle para el análisis de información

Empecemos hablando de un caso concreto de análisis descriptivo (hay otras vías), el **"Análisis Exploratorio de Datos" (EDA),** quizá más conocido como **Data Discovery.** Una vez hemos "ingerido" y almacenado un gran volumen de datos, el siguiente paso lógico sería preguntarnos cómo son esos datos y qué podremos extraer de ellos.

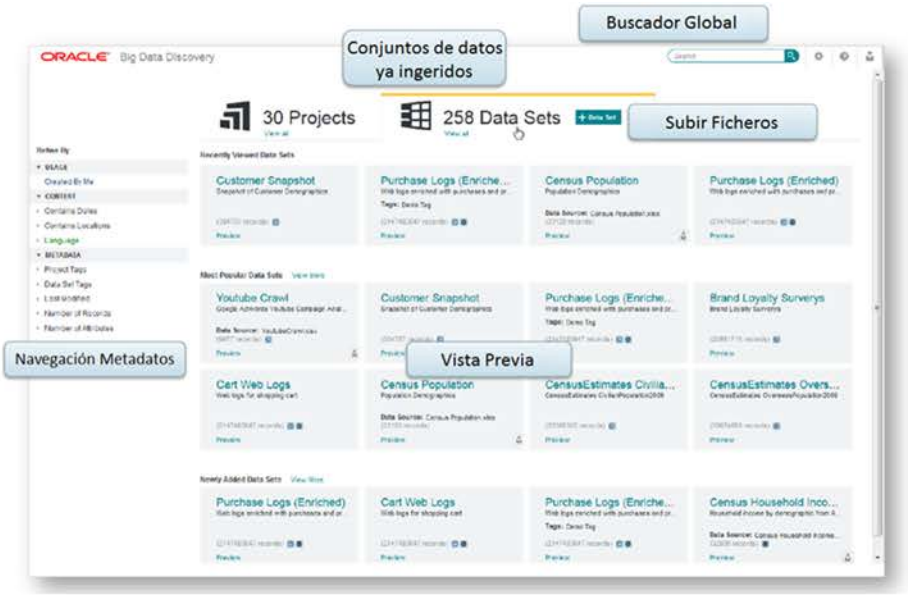
El análisis exploratorio de datos (EDA) utiliza diversas técnicas, principalmente visuales, para alcanzar varios objetivos, entre ellos:

- Detectar posibles estructuras subyacentes, patrones y/o grupos de poblaciones
- Extraer variables significativas
- Detectar datos anómalos e irregularidades
- Descartar posibles conjeturas/prejuicios preexistentes sobre un conjunto de datos

Oracle cuenta con dos herramientas destacadas para análisis exploratorio, la primera y más antigua en su portfolio, pero de la que no vamos a hablar ahora: **Oracle Endeca Information Discovery**. Aunque de alguna manera está presente en la arquitectura interna de la herramienta de la que hablaremos en primer lugar.

Oracle Big Data Discovery (BDD) es acertadamente publicitado como la "Cara visual de Hadoop", porque en esencia, es exactamente eso: una herramienta visual para analizar cualquier conjunto de datos almacenado en Hadoop, permitiendo incluso combinarlos con datos que subamos desde nuestro equipo. Una interface de usuario visual, rica e intuitiva, con un lenguaje claro que habla de Conjuntos de datos, atributos y tipos de datos permite acceder a todos los datos almacenados en el cluster Hadoop a través de una estructura de catálogo. **BDD aísla así al usuario de la complejidad del ecosistema de productos y lenguajes que corren sobre Hadoop**, ofreciéndole la posibilidad de realizar transformaciones sin necesidad de escribir código (más allá de las fórmulas explícitas, para las que también ofrece un asistente). Las transformaciones se aplican de forma "amigable": primero sobre una muestra de datos y una vez validado el proceso, lo extiende a todo el dataset. BDD dispone también de funcionalidades para análisis de sentimiento, tareas de data quality, detección de datos anómalos, etc. así como de un potente motor de búsqueda por contenidos, autor, tags, …

BDD es a la vez una herramienta descriptiva que ofrece una rica vista previa sobre los datasets del catálogo y con una completa información estadística "habitual" (medias, cuartiles, desviaciones, detectar correlaciones, …) acompañada por las necesarias representaciones gráficas en diversos formatos (boxplots, mapas, …) para explicar visualmente esa información estadística.



Y todo esto sin descuidar el objetivo final: hacer llegar el valor del Big Data al negocio. Para ello, BDD cuenta con unos dashboards fáciles de construir (drag&drop), navegables, con capacidad de exportación en varios formatos, etc. y que permiten compartir fácilmente los resultados obtenidos con el resto de la organización, pero de manera securizada, como corresponde al software corporativo.

En el itinerario hacia la cumbre de la pirámide, a menudo es necesario realizar varias iteraciones que refinen un conjunto de datos inicial. Podremos llegar a producir diversos conjuntos de datos para análisis posteriores, con objetivos y necesidades diferentes, que sigan caminos dispares, conducidos probablemente por usuarios diferentes. BDD es capaz de generar nuevos datasets con los resultados obtenidos de un análisis y/o tras la aplicación de transformaciones sobre el original, retroalimentando así el Data Pool, pero también puede exportar datos a ficheros externos con formato CSV, por ejemplo.

Como hemos visto Big Data Discovery no es únicamente una herramienta de análisis descriptivo, sino que es una herramienta muy completa que permite dar respuesta a varias necesidades.

Saltamos ahora al **análisis predictivo**. Oracle cuenta desde hace tiempo también con **Essbase,** un potente motor MOLAP in-memory con funcionalidad optimizada para predicción y simulación. Los productos de la familia Hyperion están basados en Essbase, que también puede integrarse con la OBIEE suite y que es una opción en Exalytics.

Pero en este artículo vamos a centrarnos en la opción **Advanced Analytics** de la BD, que amplía el antiguo paquete de **data mining** con **Oracle Enterprise R (ORE)** para ofrecer de manera sencilla y rápida capacidad de predicción embebida en la BD.

La aplicación de los algoritmos dentro de la BD aporta diversos beneficios:

- Eliminamos la necesidad de extraer los datos de la BD para cargarlos en una herramienta externa, y una vez obtenidos los resultados, exportarlos desde dicha herramienta y volver cargarlos en la BD (además de ahorrar unos tiempos que sumados pueden llegar a ser no menospreciables, eliminamos posibles factores de error en el proceso global)

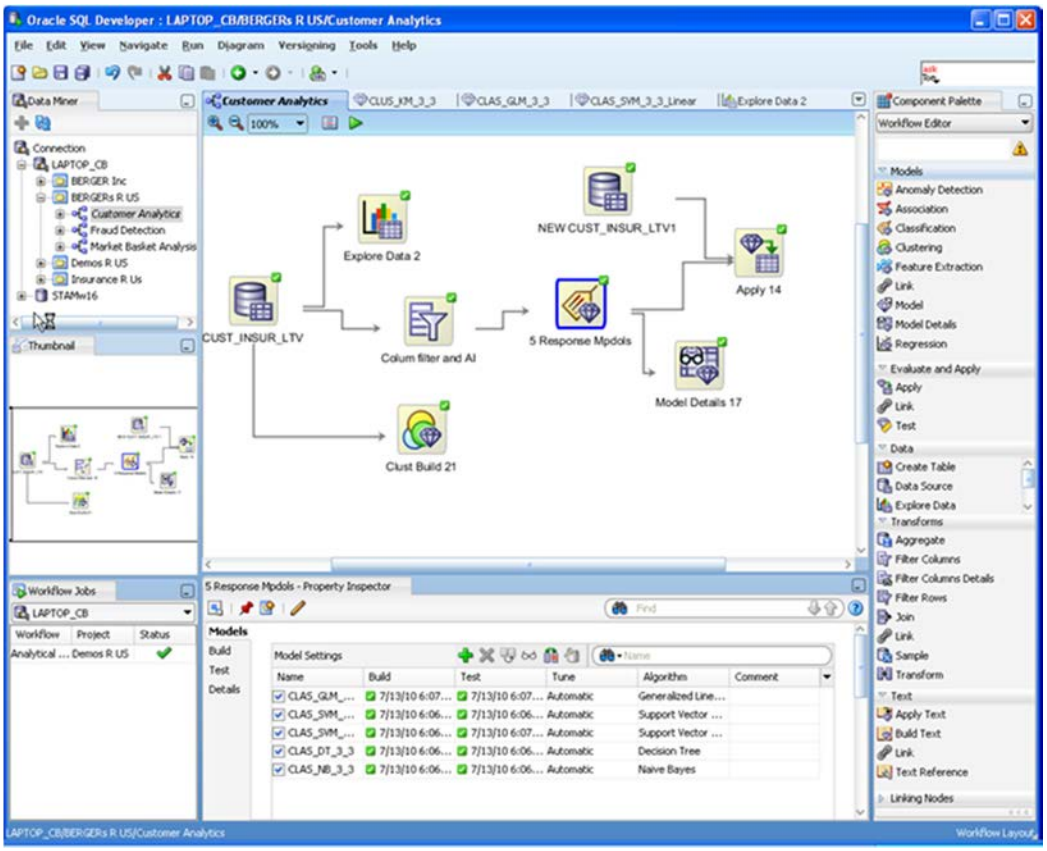
- Al estar dentro de la BD y ofrecer Oracle la funcionalidad vía SQL, automáticamente se extiende la capacidad de análisis predictivo a todas las aplicaciones que se conectan a la BD.

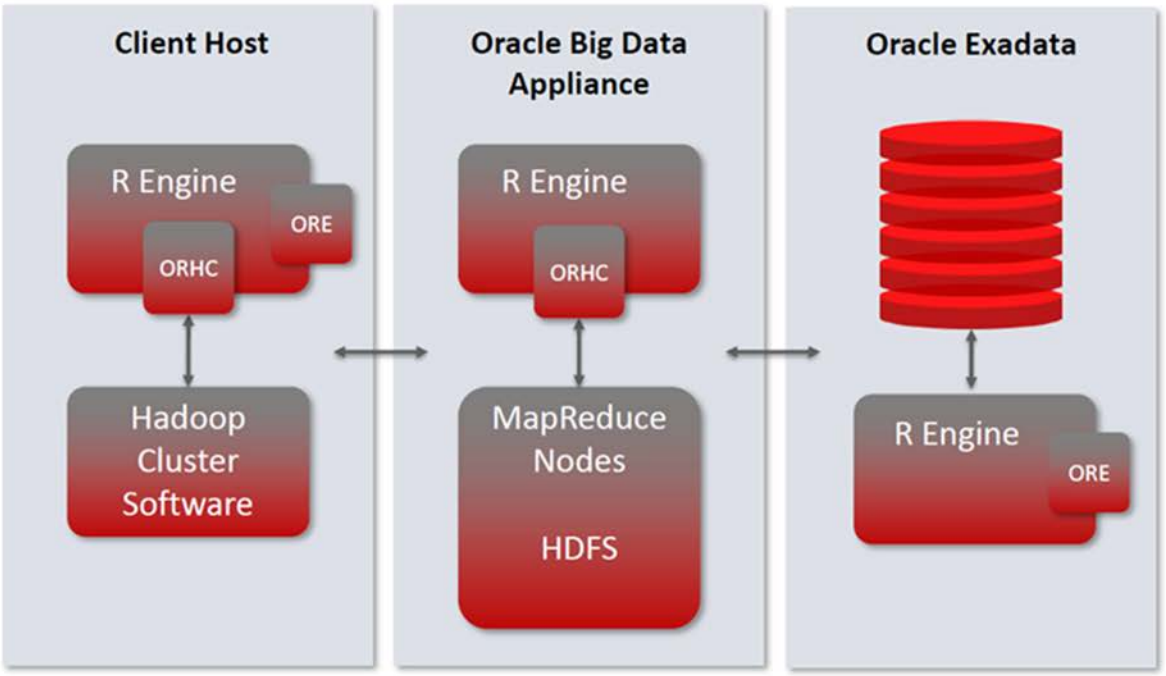
- Rendimiento y escalabilidad. Típicamente, estos procesos consumidores de recursos, quedan limitados por la potencia de cálculo y RAM disponible en el cliente que ejecuta la aplicación, mientras que al estar en BD cuentan con toda la potencia y recursos del servidor.

En la tabla siguiente, vemos un resumen de las posibilidades que nos ofrece Oracle Data Mining (ODM).

Problem	Algorithms	Applicability
Classification	Logistic Regression (GLM) Decision Trees Naïve Bayes Support Vector Machine	Classical statistical technique Popular / Rules / transparency Embedded app Wide / narrow data / text
Regression	Multiple Regression (GLM) Support Vector Machine	Classical statistical technique Wide / narrow data / text
Anomaly Detection	One Class SVM	Lack examples of target field
Attribute Importance	Minimum Description Length (MDL)	Attribute reduction Identify useful data Reduce data noise
Association Rules	Apriori	Market basket analysis Link analysis
Clustering	Hierarchical K-Means Hierarchical O-Cluster	Product grouping Text mining Gene and protein analysis
Feature Extraction	Nonnegative Matrix Factorization	Text analysis Feature reduction

ODM cuenta además con una interface de usuario gráfica y fácil de utilizar que está disponible como plug -in de Oracle SQL-Developer (que es gratuito), y con el que podremos generar flujos que automaticen nuestros procesos analíticos facilitando su despliegue y compartirlos.





La implantación de ODM y su integración es fácil y rápida, pero sus algoritmos no son ampliables o personalizables, por lo que si necesitamos expandir la funcionalidad y/o generar nuestros propios modelos predictivos, etc., podemos pensar en R.

Oracle ofrece dos posibilidades para trabajar con R a nivel corporativo: desde la BD con ORE, o desde Hadoop con Oracle R Advanced Analytics for Hadoop (ORAAH) y su conector (ORHC).

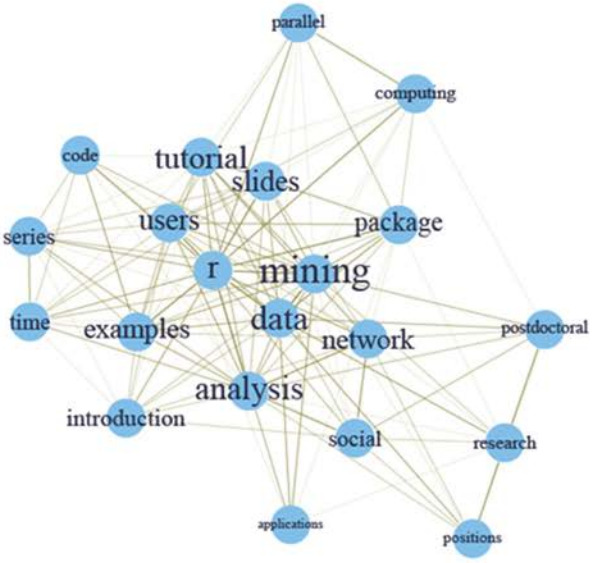
Con una filosofía análoga la seguida con ODM, ORE y ORAAH permiten acceder desde R a los datos almacenados en Hadoop o en BD, y a la inversa para almacenar los resultados obtenidos en R, logrando los mismos beneficios en cuanto a reducción de tiempo y disponibilidad de recursos, que a su vez, pueden ser puestos a disposición de todas las aplicaciones con acceso a la BD.

Volviendo al itinerario ascendente en la pirámide de valor del Big Data, debemos comentar que R es un producto interesante porqué hay una amplia comunidad desarrollando nuevos paquetes y es más fácil encontrar técnicos con conocimientos suficientes que de otros paquetes estadísticos.

Es fácil aplicar sus modelos estadísticos a problemas de negocio como por ejemplo:

- Regresión lineal para predecir la producción en función de variables de carácter meteorológico
- Regresión logística para predicciones de tipo binario como la evaluación de riesgo (devolución de un préstamo)
- Diagramas de CART y random forests cuando la predicción depende de la clasificación, ayudando por ejemplo en la prevención y aplicación de atenciones específicas grupos de población
- Clusters aplicados a clasificación de clientes (viajero ocasional, viajero habitual low-cost, …) y formulación de recomendaciones (si te gustó este artículo también de va a gustar este otro)

Con R también podemos ampliar la funcionalidad ofrecida por ODM, por ejemplo con análisis de texto para el tratamiento y clasificación masiva de correos electrónicos, detectar la posible manipulación fraudulenta de documentos, o incluso análisis de sentimiento. Otra ampliación de funcionalidad posible sería utilizar grafos para analizar redes sociales e identificar a los individuos más influyentes o bien relacionados, pequeños grupos aislados,... (como el ejemplo tomado de http://www.rdatamining.com)



Podríamos incluso avanzar un paso hacia el **análisis prescriptivo** y utilizar R para resolver problemas de optimización. Por ejemplo, en el caso de una empresa industrial, podríamos determinar la combinación óptima de productos a fabricar y/o a qué clientes servir en primer lugar, en un período de tiempo determinado y en base a la demanda existente, los costes de producción y las restricciones de recursos, de manera que se obtenga el máximo beneficio.

Aunque en el portfolio de Oracle ya existen productos estándar de análisis prescriptivo.

Oracle Real-Time Decisions (RTD) es un producto diseñado para optimizar los procesos de negocio apoyándose en diversas técnicas analíticas, como reglas de negocio, data mining, modelos estadísticos y aprendizaje automático (Machine Based Learning).

RTD aplica dichas técnicas sobre la información que logra obtener, para determinada entidad, a través de todos los canales a su alcance:

- Comercio electrónico
- Call Centers
- Tiendas
- Interacción en las redes sociales
- Flujos de datos internos

RTD utiliza modelos estadísticos en tiempo real, siendo capaz de adaptarlos a medida que detecta cambios en el comportamiento del usuario y correlacionando automáticamente centenares de atributos, que aplica específicamente a cada objetivo.

En base a los objetivos de negocio definidos para cada proceso concreto y los inputs de cada operación, y aplicando modelos que corresponden a diversas perspectivas, como canales de interés, ofertas en las que se ha interesado, criterios de retención de clientes según su edad, sexo, etc. o las interacciones que ha abandonado, **RTD es capaz de sugerir en tiempo real, la mejor acción a realizar por parte de nuestro negocio**, como por ejemplo, predicciones y/o recomendaciones personalizadas sobre los productos que pueden interesar a un consumidor.

La plataforma es sólo un medio

La reflexión tras esta exposición es que para **extraer todo el potencial del Big Data, las organizaciones deben ser capaces de realizar Advanced Analytics**; no es suficiente realizar un proyecto Big Data, que se queda en las capa de infraestructura, para extraer su valor.

Aunque sería falso decir que no obtendríamos información con él, estaríamos renunciando a la parte más novedosa y ventajosa, el análisis predictivo y prescriptivo, tanto sobre datos estructurados como no estructurados (ahora posible, gracias al Big Data).

Desde nuestro punto de vista, a medida que vayamos ascendiendo por la pirámide de conocimiento del Big Data, iremos incrementando el valor obtenido: transformando los datos en información, de la que extraeremos conocimiento, que nos permitirá tomar decisiones más acertadas.

Dicho de otra manera: superada la base de la pirámide donde residen los datos mayoritariamente crudos (que de alguna manera capturamos y persistimos…), y a través del **análisis descriptivo** lograremos identificar nuevas variables, a las que podremos asignar un valor (si no a todas, al menos sí a muchas) y/o asociarlas a ocurrencias de algunas de las entidades que ya existen en nuestro DWH (clientes, artículos,…) enriqueciéndolas. Seguidamente, podremos combinar con un **análisis diagnóstico** que a su vez genere datos para nuevas variables (que no serán hipótesis, sino reales) y con los que podremos construir nuevos datasets (o enriquecer de nuevo los existentes).

Sobre esta información (que no teníamos), podremos aplicar modelos estadísticos existentes o desarrollar nuevos modelos y realizar **análisis predictivos** de diversos tipos según el objetivo que fijemos, con elevada fiabilidad (pues se basan en datos reales) para tomar decisiones más acertadas… o incluso permitir que un sistema de **análisis prescriptivo** automatice la toma de decisiones.

La conclusión es que **ambos conceptos deben ir de la mano: que Big Data no es una finalidad, sino un medio para llegar a Advanced Analytics.**

Cuanto "más alto" logremos llegar, mejor, pero se trata de una ascensión, por lo que cuanto más fácil sea el camino, mejor. Tomar un camino difícil (curva de aprendizaje lenta: muchos lenguajes y herramientas nuevas a aprender e integrar), equivocado (sin soporte, riesgo de discontinuación) o poco transitado (con pocos técnicos cualificados, tecnologías de nicho) alargará el tiempo necesario para que el proyecto de Big Data empiece a dar frutos e incrementará los costes de desarrollo.

Por eso invertir en productos Oracle para desarrollar proyectos Big Data + Advanced Analytics en lugar de elegir otros caminos, probablemente termine suponiendo un ahorro. Sí, un ahorro en términos de time to market y de optimización de recursos que deberemos destinar a la selección de herramientas, despliegue e integración, en el momento inicial. Y a medio y largo plazo, un ahorro en horas de formación, desarrollo y búsqueda de soporte, durante el ciclo de mantenimiento del proyecto.