CÓMO SE MANEJAN LOS GRANDES VOLÚMENES DE DATOS

AUTOR: GASTÓN ADDATI

CONTENIDO

INTRODUCCIÓN

El big data surge como consecuencia del aumento exponencial de los datos. Sin dudas, el surgimiento de internet y redes sociales han aumentado considerablemente los datos que se generan y se comparten en todo el mundo. Las preguntas que surgen son las siguientes: *¿cómo se puede lograr el procesamiento de estos grandes volúmenes de datos? ¿Qué tecnologías necesitaremos implementar en proyectos de big data? ¿Cómo se realiza dicha implementación?*

Durante este libro pondremos especial atención en analizar cómo funciona el big data, considerando las fuentes diversas y heterogéneas desde donde realizaremos la captura (o ingesta) de los datos, el procesamiento para convertir dichos datos en información y desde luego, como se hace para transformar esa información en conocimiento.

VIDEO

¿Qué es BIG DATA y para qué sirve?

01

¿CÓMO SE MANEJAN LOS GRANDES VOLÚMENES DE DATOS?

¿Cómo se manejan los grandes volúmenes de datos? Este es el gran interrogante que existe y sobre el cual pondremos especial atención.

Los proyectos de big data intentan realizar un proceso de incorporación masiva de datos con el objetivo de transformarlos en información que aporte valor a una organización. Este proceso de transformación implica cuatro grandes partes, que son las que dan explicación completa de cómo funciona el big data. El proceso genérico de manejo y transformación de datos:

1. Capturar los datos
2. Almacenar y Procesar los datos
3. Analizar los Datos
4. Puesta en Valor

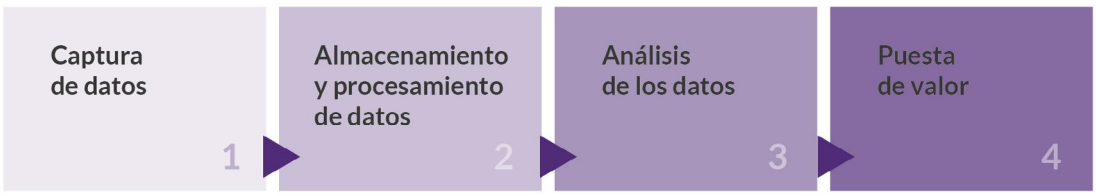


Figura 1. Proceso de manejo de volúmenes de datos (Elaboración propia).

A continuación, realizaremos una descripción de cada uno de estos procesos.

1.1. CAPTURA DE DATOS

Durante esta fase, el proceso se centrará en comprender y dar respuestas a ciertas preguntas tales como: *¿Dónde se encuentran los datos que voy a recopilar? ¿Cómo podremos capturarla? ¿Qué tipo de datos recopilaré?. ¿Con qué frecuencia los recopilaré?* Este es uno de los primeros pasos que realizaremos en los proyectos de big data, porque como sabemos, hay mucha información dispersa y las fuentes serán múltiples (redes sociales, páginas web, otros sistemas de información, etc.). También sabemos que los datos que se capturarán podrán ser datos estructurados, no estructurados o, incluso, semi-estructurados (datos de todo tipo). Típicamente, los proyectos de big data incorporan datos como emails de clientes o posibles clientes, publicaciones en redes sociales, podcasts, videos, documentos PDF, entre otros. Este proceso de incorporación de datos también se denomina *"ingesta"* de datos donde, además de conocer a dónde ir a buscarlos, deberemos identificar con qué frecuencia realizaremos la ingesta, ya que no es lo mismo recopilar datos una vez al día, a la semana o al mes, que realizar una ingesta de manera constante en tiempo real o casi real. Si este fuera el caso, estaremos hablando de proyectos de big data streaming.

1.2. ALMACENAMIENTO Y PROCESAMIENTO DE DATOS

Durante este proceso, teniendo mejor claridad, sobre todo por haber pasado por el proceso previo, nos ocuparemos de analizar qué haremos con los datos capturados. La pregunta natural en esta etapa es: *¿qué hacemos con los datos? ¿Dónde los almacenaremos?* Este proceso tiene relación directa con lo que analizaremos luego en relación

NOTAS

El procesamiento de datos de streaming resulta beneficioso en la mayoría de las situaciones en las que se generan datos nuevos y dinámicos de forma continua. Es apto para la mayoría de los sectores y casos de uso de big data.

con la definición de una arquitectura de big data, porque será importante definir qué tipo de tecnología de base de datos es la mejor opción, y sobre todo cual es la recomendada para almacenar los datos que necesitamos.

También debemos considerar en esta etapa que los datos deberán ser procesados. Ese procesamiento puede realizarse en tiempo real (*streaming*) o a demanda o cada cierto tiempo (se denomina proceso *batch*).

1.3. ANÁLISIS DE LOS DATOS

La fase del análisis de los datos es tal vez una de las más ricas, interesantes y complejas a la vez en todo proyecto de big data, porque los datos que ya almacenamos requieren de un correcto análisis, de la aplicación de diversos métodos complejos para identificar patrones, comportamientos y segmentar la información. En esta etapa es natural que se implementen algoritmos de inteligencia artificial para construir y modelar situaciones que le permitirán a la organización realizar análisis del tipo descriptivo (poniendo énfasis en lo que ya ocurrió en el pasado), del tipo predictivo (identificar tendencias y armar pronósticos hacia el futuro) o bien, del tipo prescriptivo (sugerencias de qué debe hacer la organización y cómo hacerlo).

1.4. PUESTA EN VALOR

El último de los procesos también es importante y necesario. Sin este proceso, todo lo anterior carece de sentido. El objetivo principal de la puesta en valor es presentar los datos convertidos en información. Esta información debe servir y debe acompañar la toma de decisión dentro de una organización. Las decisiones que tomará la organización

IMPORTANTE

La definición de qué tipo de procesamiento tendremos que realizar es crucial en esta etapa, porque si el proyecto es de streaming (captura y procesamiento en tiempo real o casi real) cambia considerablemente (desde el punto de vista de la tecnología y la implementación) a que si el procesamiento es batch.

estarán sustentadas y respaldadas por datos (los cuales consideramos fidedignos). El objetivo que persigue la puesta en valor es crucial, porque como sabemos, los datos por si solos, no nos dicen nada. Debemos descubrir, comprender y agilizar nuestra organización para que las decisiones que se tomen (en línea con el negocio) sean las mejores y sobre todo estén basadas en datos.

Es con este último proceso donde las organizaciones pueden generar valor agregado a sus clientes, donde las organizaciones pueden ser más competitivas, mejorar su calidad de atención, mejorar sus productos, descubrir nuevos servicios, entre otras tantas cuestiones donde el fin será la generación de ventaja competitiva.

02

ANTES DE LA TECNOLOGÍA: CONOCER EL NEGOCIO

Uno de los principales errores que comenten las organizaciones al momento de implementar cualquier tipo de solución tecnológica es pensar en la implementación desde el punto de vista de la tecnología (generalmente, la más conocida) de un software, o de un hardware o de un servicio tecnológico, sin antes analizar y entender del todo cual es el problema de negocio que debo resolver.

Tal como lo plantean Kenneth & Jane Laudon en su libro Sistemas de información gerencial, *"La introducción de un nuevo sistema de información implica mucho más que un nuevo hardware y software. También implica cambios en los trabajos, habilidades, administración y organización [...] Crear un nuevo sistema implica un cambio organizacional planeado"*.

Algunas herramientas que se pueden utilizar para esto son las siguientes: Análisis o Matriz FODA, Diagramas de flujo, Diagramas Causa-Efecto, Diagramas para el armado de modelo de negocio (del estilo CANVAS), o cualquier otra herramienta gráfica o explicativa que permita plasmar la problemática y que sirva de guía para el proyecto de big data.

IMPORTANTE

La recomendación en este sentido que se le puede brindar al lector es que, antes de comenzar un proyecto de big data, se analice, se conozca y se comprenda con la mayor precisión posible el problema de negocio que el proyecto de big data pretende resolver. Esto es sumamente importante, porque cuanta más información se cuente en este sentido, mejor acompañada estará la estrategia tecnológica que se tomará para resolver el problema planteado.

Otra recomendación o sugerencia es que se pueda pensar al proyecto de big data, identificando a los principales interesados (o stakeholders) del proyecto, y que al momento de proponer objetivos a cumplir o cuestiones de negocio a resolver, estos objetivos o cuestiones sean realistas. Ocurre que muchas veces se plantean objetivos muy ambiciosos, difíciles de cumplir y por supuesto, inviables de implementar. Sea realista, concreto y piense que es un proyecto a largo plazo, implementado por fases o etapas, asegurando que cuenta con el respaldo de la organización.

A partir de un buen diagnóstico, de un buen análisis de situación actual, de la fijación de objetivos claros, medibles y cuantificables, recién en ese momento, estará en condiciones de pasar a la siguiente fase.

NOTAS

Según el PMI (Project Management Institute), los Stakeholders son cualquier persona u organización cuyos intereses pueden afectar, de manera positiva o negativa, la realización del proyecto.

03

ORIGEN Y TIPOS DE FUENTES DE DATOS

Definido el problema a resolver, y los objetivos a cumplir, llega el momento de pensar lo siguiente: *¿Qué tipo de datos necesito para resolver el problema o cumplir los objetivos? ¿Dónde están los datos? ¿Cuáles son las fuentes de información a las que puedo recurrir?*

Estas y otras preguntas son las que se deben intentar responder en esta etapa, en la cual, aún no tenemos datos, ya que solo los estamos pensando desde un punto de vista lógico y conceptual.

Generalmente ocurre que los datos que deseamos obtener son datos que fácilmente podríamos conseguirlos (porque están dentro de nuestra organización o porque podemos acudir a otros sistemas para obtenerlos). Pero también ocurre que, muchas veces, esos datos no son factibles de encontrar o de utilizar, o bien podrían serlos, pero habría que realizar programas que capturen la información de manera automática (con técnicas como *web scraping*, por ejemplo).

NOTAS

La técnica del web scraping es muy utilizada para obtener información de manera automatizada. Mediante un programa informático y utilizando esta técnica, se puede analizar y capturar cierta información requerida por el usuario, desde diversas páginas web.

Durante esta fase, es fundamental el trabajo que realiza el **Analista de Negocios o el Business Analyst,** quien es la persona que debe poseer una visión amplia sobre el negocio, los procesos y la tecnología. Este analista, entre tantas otras cosas, clasificará las fuentes de datos (sobre todo en fuentes internas y externas).

Las fuentes internas son aquellas que existen dentro del ámbito organizacional. Es decir, aquellas fuentes dentro de la organización que son capaces de brindarnos la información que deseamos. Típicamente estas fuentes internas provienen de sistemas de información como pueden ser sistemas CRMs, ERPs, Sistemas de Gestión de Operaciones, Sistema de cadenas de suministros, por mencionar algunos ejemplos. Diversos sistemas y bases de datos dentro de la organización se transforman en fuentes internas, incluso aquellas planillas de cálculo que contienen información de clientes y que en empresas pequeñas es muy frecuente encontrar.

Las fuentes externas, en cambio, son aquellas que no se encuentran o se consiguen dentro de la organización. Son fuentes de datos que pueden entregar otros sistemas informáticos, otros sistemas accesibles en la web (como redes sociales), otros sistemas de clientes, de proveedores, o incluso otras bases de datos que se pueden comercializar, y en general son fáciles de identificar porque se encuentren fronteras fuera de la organización. En conclusión, toda la información que no esté dentro de la organización es una fuente externa.

El analista de negocios culminará esta etapa haciendo un trabajo muy complejo de análisis de fuentes de datos y de información. Para eso utilizará diagramas específicos que ayudarán a entender la relación entre las diversas entidades que existen, y sobre todo, a entender la relación entre toda la información que se desea recopilar, para aportar valor agregado al negocio.

A continuación, un ejemplo de diagrama básico para darnos una idea de cómo se generan las relaciones de los datos, que ayudan al entendimiento.

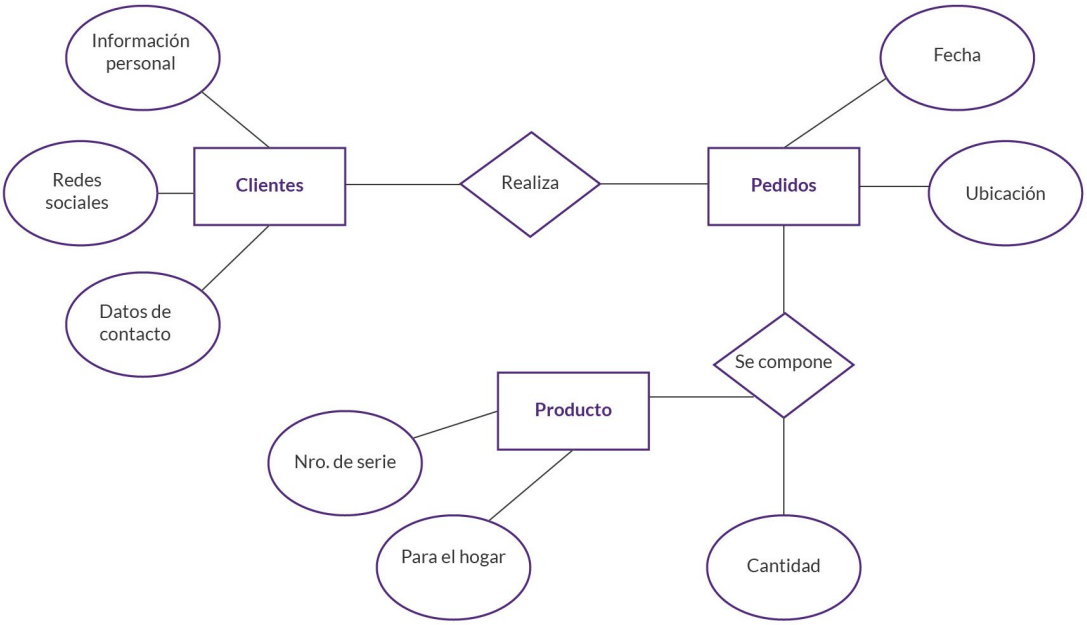


Figura 2. Ejemplo de diagrama (Elaboración propia).

IMPORTANTE

Comprender los datos es crucial para cualquier proyecto de big data.

04

DISEÑO DE UNA ARQUITECTURA TECNOLÓGICA QUE DÉ SOPORTE AL NEGOCIO

Una de las cuestiones más importantes en todo proyecto de big data es la cuestión vinculada a la arquitectura tecnológica que deberá diseñarse e implementarse para dar respuestas al problema planteado y al cumplimiento de los objetivos propuestos.

Una arquitectura se puede representar como un entorno actual y futuro de nuestro negocio, donde se describe de manera general cada aspecto del negocio y de la tecnología, donde se desarrolla en una serie de capas hasta llegar al nivel más bajo posible de entendimiento.

Una arquitectura, entonces, es una descripción formal de un sistema de información, organizada de manera que permite el razonamiento acerca de sus propiedades estructurales.

CONCEPTO

Una arquitectura tecnológica es un modelo conceptual que define la estructura, el comportamiento, la gobernabilidad y las relaciones entre el hardware, software, las redes, los datos y la interacción humana más allá de todo el ecosistema que rodea los procesos de negocios.

Desde el punto de vista práctico, existen muchos modelos de referencia sobre los cuales los principales consultores o ingenieros se apoyan. Lo importante de las arquitecturas de referencia es que sirven para modelar y diseñar la arquitectura que mejor se adapta a las necesidades o problemas del negocio según la industria en cuestión.

Veamos un ejemplo de un modelo de arquitectura para Big data:

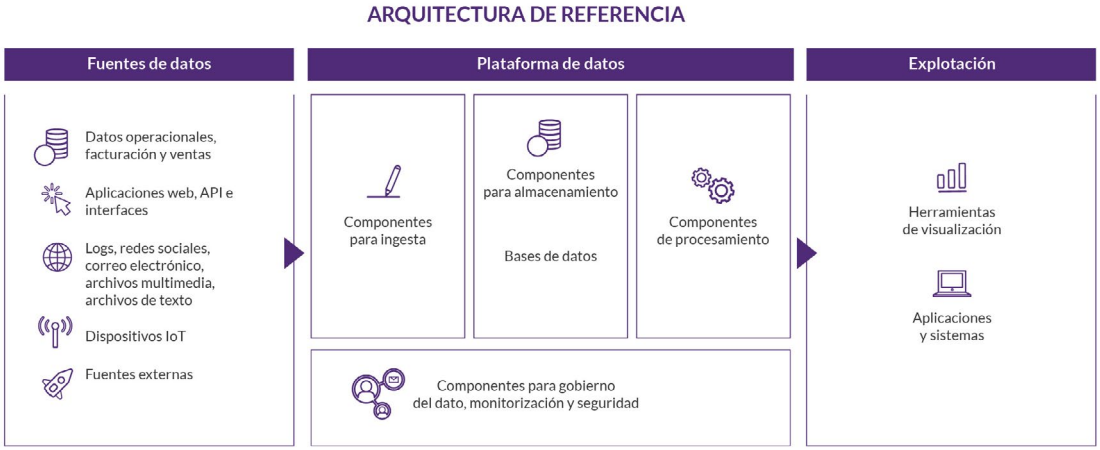


Figura 3. Arquitectura de referencia para big data (Jesús Montoya Sanchez de Pablo).

En lo que respecta a las fuentes de datos, esta es la capa que representa, identifica y sitúa los diversos sistemas, aplicativos, bases de datos y otras fuentes que constituyen el origen de los datos que se van a recopilar. Recordemos que en las arquitecturas de para big data será factible incorporar cualquier tipo de dato, con independencia de su naturaleza (estructurado, no estructurado o semiestructurado).

La capa que corresponde a las plataformas de datos se constituye de diversos componentes que conllevan a definir todo lo relacionado a la lógica del procesamiento de los datos. Desde cómo se realizará la ingesta (*¿en tiempo real o en un proceso Batch?*). También se definirá la tecnología de base de datos a utilizar, donde típicamente en proyectos de big data y dada su naturaleza de tratamiento de datos diversos, se emplean tecnologías NoSQL (como veremos luego, es muy utilizada la tecnología MongoDB o Cassandra). También en esta capa se analiza y se implementan algoritmos de procesamiento de datos basados en inteligencia artificial, y también se especifican cuestiones vinculadas al gobierno de los datos y a la seguridad.

NOTAS

La gestión de datos consta de políticas, procesos y una estructura organizativa para dar soporte al control de datos empresariales. La estructura de un programa de gobierno de datos proporciona comprensión, seguridad y confianza en torno a los datos de una organización y a sus stakeholders, especialmente a medida que las empresas escalan y acumulan más activos y fuentes de datos.

Por último, la capa de explotación de los datos: aquí es donde se definen las diversas herramientas de visualización, especificación de reportes, análisis estadístico, y todo lo referido a mostrar de una manera adecuada los resultados obtenidos luego de todo el modelado, procesamiento y análisis de la información.

Las arquitecturas son muy importantes porque permiten no solo resolver el problema planteado, sino que deben estar preparadas para el futuro crecimiento tanto del negocio, como de los datos. Por eso es sumamente recomendable que las arquitecturas elegidas puedan ser versátiles y tener escalabilidad.

Una de las cuestiones trascendentales que habrá que definir al momento de diseñar la arquitectura es si trabajaremos bajo el esquema o un modelo OnPremise, es decir, donde toda la infraestructura tecnológica de servidores, aplicaciones, seguridad y almacenamiento residirá en los datacenters de la organización, o si todo, o parte de eso, se alojará en la nube (Modelo Cloud), o incluso si trabajaremos bajo un modelo híbrido, es decir, una parte OnPremise (almacenamiento y procesamiento de datos) y otra Parte (por ejemplo el Analytics) en Cloud.

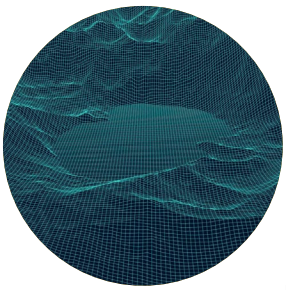
05

EL TRATAMIENTO DE LOS DATOS

5.1. FASE DE PREPARACIÓN

La elección y puesta en marcha de una arquitectura robusta nos permitirá trabajar con los datos y, sobre todo, procesarlos de una manera correcta para obtener información.

Durante esta fase, la atención estará puesta en cómo se realizará el proceso de captura (ingesta) y el proceso de almacenamiento para que los datos con los que vayamos a trabajar estén en condiciones de ser procesados. A todo este proceso se lo denomina técnicamente *"tratamiento de los datos"* y, si bien su nombre parece algo simple, en la práctica implica cuestiones informáticas muy complejas que llevan adelante un equipo de profesionales técnicos, expertos en tecnología y expertos en la solución de big data que se esté implementando.



Cuando se habla de big data, se habla de **data lakes** (lagos de datos). Este concepto es sumamente importante para la explicación de esta fase. Vamos a profundizar en este concepto.

En líneas generales, los datos que vamos a recopilar (estructurados o no estructurados) requieren de un cierto ordenamiento y tratamiento para poder dejarlos listos para un posterior tratamiento.

También será importante en esta instancia tener definiciones claras y concretas acerca de cómo voy a capturar esa información y cada cuanto tiempo. Recordemos que una de las principales cuestiones que van a caracterizar a la ingesta de los datos es saber si se trata de un proyecto *Streaming* o de un proyecto donde a los datos pueda buscarlos en un proceso Batch. Dependiendo de uno o de otro, es la tecnología de ingesta que utilizaré (lo veremos luego).

Lo que suelen hacer los especialistas en big data, con toda esta información, es diseñar los denominados **data lakes**, que son repositorios donde van a almacenarse todo tipo de datos y donde, además, se van a procesar para transformar y extraer los datos para su posterior análisis y visualización.

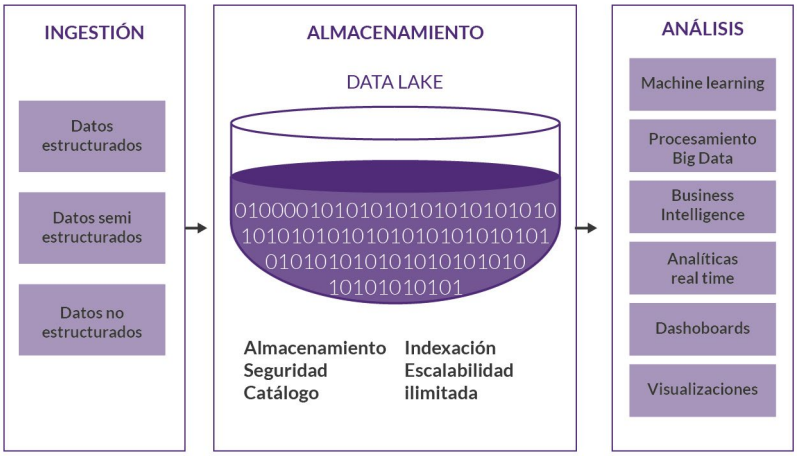


Figura 4. Data lake (Conceptos clave sobre Data Lakes).

Un data lake es una modalidad nueva y cada vez más popular de almacenar y analizar datos, porque permite a las empresas administrar múltiples tipos de datos de una amplia variedad de fuentes, y almacenar estos datos, estructurados y no estructurados, en un repositorio centralizado.

CONCEPTO

Los lagos de datos o data lakes son repositorios donde van a almacenarse todo tipo de datos y donde, además, se van a procesar para transformar y extraer los datos para su posterior análisis y visualización.

5.2. FASE DE MODELIZACIÓN Y VISUALIZACIÓN

Una vez que los datos ya están disponibles en el data lake, nos va a interesar trabajar con ellos, fundamentalmente modelándolos de cierta forma y con ciertas tecnologías, para que, a partir de esos modelos, podamos obtener y visualizar información.

En esta etapa de modelización es donde típicamente se implementa la inteligencia artificial y sobre todo el Machine Learning (como una rama de esta).

> El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas (del inglés, machine learning) es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan.

Dentro de lo que se conoce fundamentalmente como machine learning, trabajaremos con dos tipos de modelos de aprendizajes: el modelo de aprendizaje supervisado y el modelo de aprendizaje no supervisado:

1. **Modelos de aprendizaje no supervisado.** Estos algoritmos de aprendizaje se utilizan para agrupar los datos no estructurados según sus similitudes y patrones distintos en el conjunto de datos. El término *"no supervisado"* se refiere al hecho de que el algoritmo no está guiado como sí lo está el algoritmo de aprendizaje supervisado. Los algoritmos de aprendizaje no supervisados permiten realizar tareas de procesamiento más complejas en comparación con el aprendizaje supervisado. Los algoritmos de aprendizaje no supervisado manejan datos sin entrenamiento previo, es una función que hace su trabajo con los datos a su disposición. El aprendizaje no supervisado pretende descubrir patrones previamente desconocidos en los datos, pero la mayoría de las veces estos patrones

LINK

son aproximaciones deficientes de lo que el aprendizaje supervisado puede lograr.

2. **Modelos de aprendizaje supervisados.** El aprendizaje supervisado proporciona una ruta directa para convertir datos en información real y procesable. Este aprendizaje es uno de los motores más potentes que permite que los sistemas de inteligencia artificial tomen decisiones empresariales de forma más rápida y precisa que los humanos.

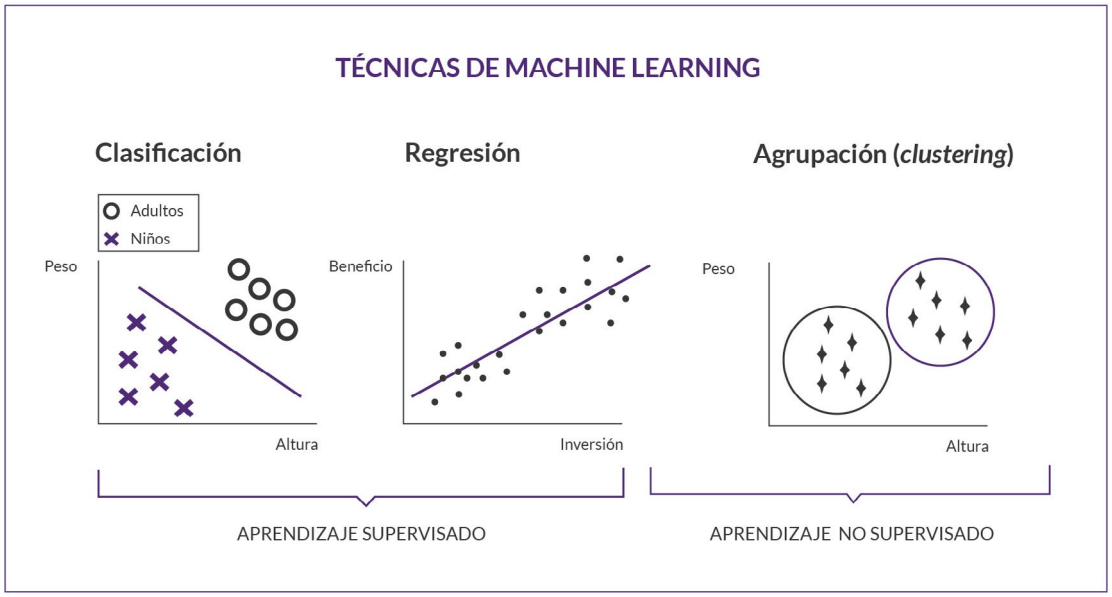


Figura 5. Técnicas de machine learning: Aprendizaje supervisado vs. aprendizaje no supervisado (Modelos de Machine Learning).

Dependiendo de cada proyecto, los datos que se hayan podido recopilar, el procesamiento de esos datos y la calidad del modelo programado con Inteligencia Artificial, tendremos resultados que pueden o no ser los esperados, pero serán resultados al fin.

La próxima medida, y luego de haber ensayado el o los modelos de machine learning, es visualizar la información obtenida. Para la visualización de la información se podrán utilizar diversas herramientas y tecnologías (las cuales fueron definidas en la arquitectura). Será crucial en esta etapa poder acompañar la visualización con una buena y acabada descripción de la información obtenida.

NOTAS

En algunas organizaciones incluso, existen roles específicos para esto. Personas entrenadas en cuestiones de diseño gráfico y en storytelling que colaboran y aportan valor en la presentación de la información. Son los denominados Data Artists.

Muchas veces ocurre que los reportes son tan elegantes que, al mismo tiempo, son poco entendibles. Será importante aquí la presencia de un analista o de una persona que pueda ser capaz de traducir con simplicidad los resultados y las conclusiones.

Para que los reportes puedan ser presentados con mayor claridad y simpleza, existen algunas herramientas básicas como las siguientes:

1. **Informes y reportes.** Una conjunción de texto, imágenes, gráficos que tienen como objetivo describir la información obtenida. A modo de ejemplo, un informe o un reporte puede tener la siguiente forma:

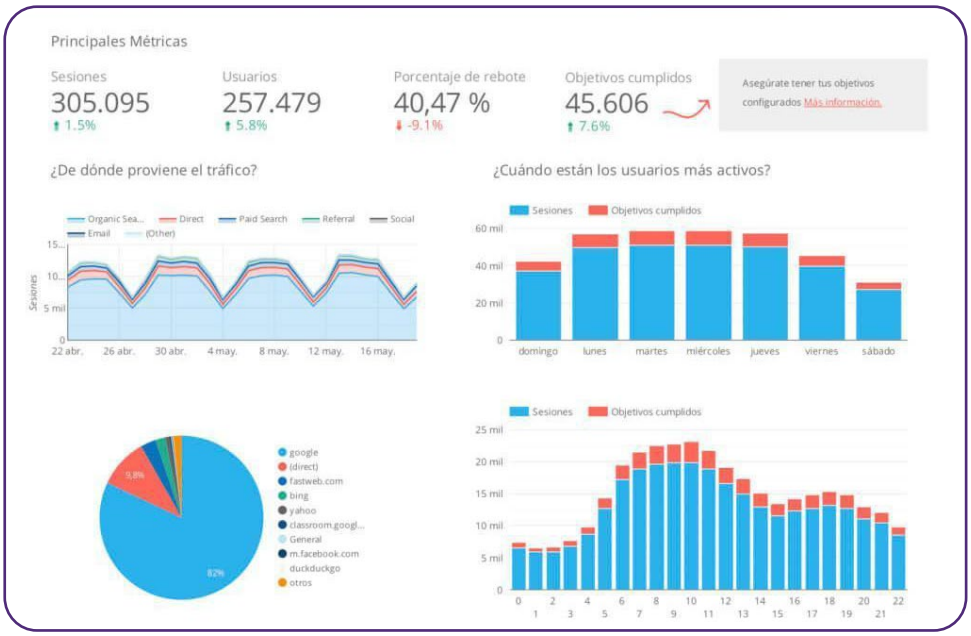


Figura 6. Ejemplo de presentación de datos

LINK

Puede visitar el siguiente sitio para interiorizarse acerca de POWER BI. Una herramienta muy utilizada en la actualidad para visualización y reportes: *https://powerbi.microsoft.com/es-es/power-bi-visuals/*

2. **Infografías interactivas.** Las infografías son similares a los reportes que mencionamos anteriormente, pero, en algunos casos, le da la facilidad al usuario de poder responder a ciertas preguntas con los datos que fueron preparados para tal fin. Incluso estas infografías pueden ser dinámicas y navegables.

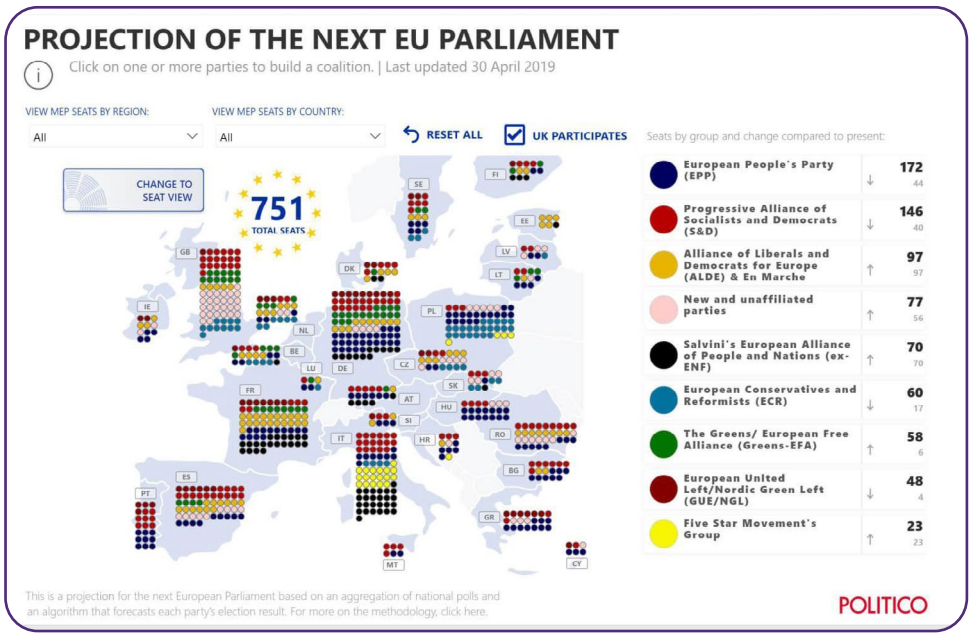


Figura 7. infografía (Objetos visuales de Power BI | Microsoft Power BI).

IMPORTANTE

Esta fase de visualización y de reporte de la información es una de las más importantes dentro de los proyectos de big data, porque permite generar conocimiento y darlo a conocer, sobre todo a las personas que tomarán las decisiones. Por tal motivo, y por ser una herramienta de apoyo a las decisiones basadas en datos, es menester que sean presentadas de una manera correcta.

5.3. FASE DE PUESTA EN VALOR

William Edwards Deming, quien fuera uno de los autores más influyentes del siglo XX en cuestiones relacionadas al concepto de calidad total, ha dicho una frase tan simple y completa que aún hoy, en tiempos del big data, aplica a la perfección: *"Sin datos, solo eres otra persona más dando su opinión"*. Esta frase de W. Deming, puesta de manifiesto antes del surgimiento del big data, refleja de alguna forma el cambio de paradigma en la forma en la que las organizaciones toman decisiones.

Algunas organizaciones toman decisiones basadas en la experiencia, o en el criterio del conductor (ejecutivo responsable), y muchas veces esa decisión es tomada sobre la base de un modelo como pueden ser el modelo racional, el modelo intuitivo, o el modelo psicológico de toma de decisiones.

Lo cierto es que el big data permite que las organizaciones dejen de tomar decisiones basadas en la intuición o en lo que hace la competencia para tomar decisiones basadas en datos, porque sin datos, y sobre todo en el contexto actual en el que vivimos (plena era digital o revolución industrial 4.0), los datos están por todos lados, nos invaden y de ellos debemos sacar provecho *¿Cómo sacaremos provecho?* Sin duda, utilizando el big data y la Inteligencia Artificial. Teniendo información certera, que permita aportar conocimiento no sólo de lo que ocurrió en una organización determinada, sino aportar posibles tendencias, comportamientos, en fin, hacer predicciones (análisis predictivo y/o prescriptivo).

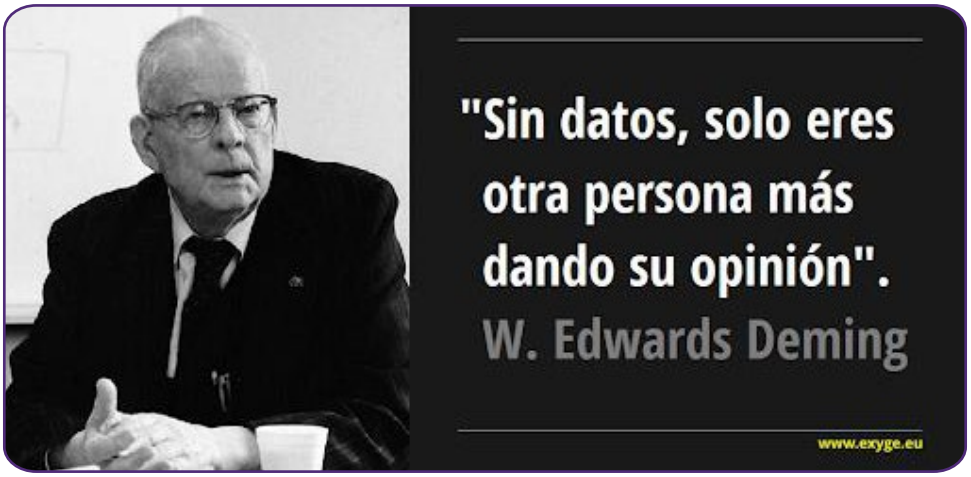


Figura 8. Cita Deming (www.exyge.eu).

Por otro lado, también se habla de puesta en valor para ciertos modelos o ciertos casos de uso que una organización implementa. Por ejemplo, un banco podría tener un modelo entrenado de inteligencia artificial para detectar si un cliente puede acceder a un cierto beneficio (upgrade o actualización de cuenta bancaria) en función del promedio de ingresos que se registraron en el último mes. O también el otorgamiento de créditos automáticos sobre la base de la información financiera de una persona o de una empresa.

Otro ejemplo puede ser la actualización o la renovación de una póliza de seguro, en el momento y sin intervención humana donde el modelo puede, en función de los datos de la persona, su historia y su capacidad de conducir, otorgar premios o castigos al momento de renovar o adquirir una nueva póliza.

Por supuesto, y dado que los modelos son autónomos, en estos casos se requiere mucho seguimiento y control de las decisiones, no sólo para ajustar los modelos, si no para medir y cuantificar los resultados obtenidos.

La puesta en valor requiere un proceso continuo y sistemático de revisión permanente entre lo que hacemos, lo que decidimos y lo que cambiamos como organización.

06

HADOOP, EL PADRE DEL BIG DATA

Muchos autores y fuentes de información acuerdan en que Hadoop es el padre del big data, lo cual en cierto punto es cierto, pero no podemos dejar de mencionar que la empresa Google, a inicios del año 2000, comenzó a investigar nuevas formas para tratar los grandes volúmenes de datos, el almacenamiento distribuido y nuevos métodos de acceso a la información.

Hubo 3(tres) artículos académicos (científicos) que dan prueba de estos avances:

1. *"The Google File System (GFS)"*. Artículo publicado en octubre 2003 (revista de ACM).
2. *"MapReduce: Simplified Data Processing on Large Clusters"*. Artículo publicado en diciembre de 2004, en OSDI'04.
3. *"Big Table: A Distributed Storage System for Structured Data"*. Artículo publicado en noviembre de 2006, en el OSDI'06.

Podríamos decir que el mundo comenzaba a conocer tres grandes elementos esenciales: *Google File System (GFS), MapReduce y Big Table.*

CONCEPTO

Hadoop es un entorno de trabajo para software, bajo licencia libre, para programar aplicaciones distribuidas que manejen grandes volúmenes de datos (big data). Permite a las aplicaciones trabajar con miles de nodos en red y petabytes de datos. Hadoop se inspiró en los documentos de Google sobre MapReduce y Google File System (GFS).

Hadoop es una implementación open source (de código abierto) de **MapReduce** que fue fundada en sus orígenes por la empresa Yahoo! en el año 2006. La idea original era, justamente, trabajar con grandes volúmenes de datos haciendo más efectivo el acceso a la información y optimizando el almacenamiento.

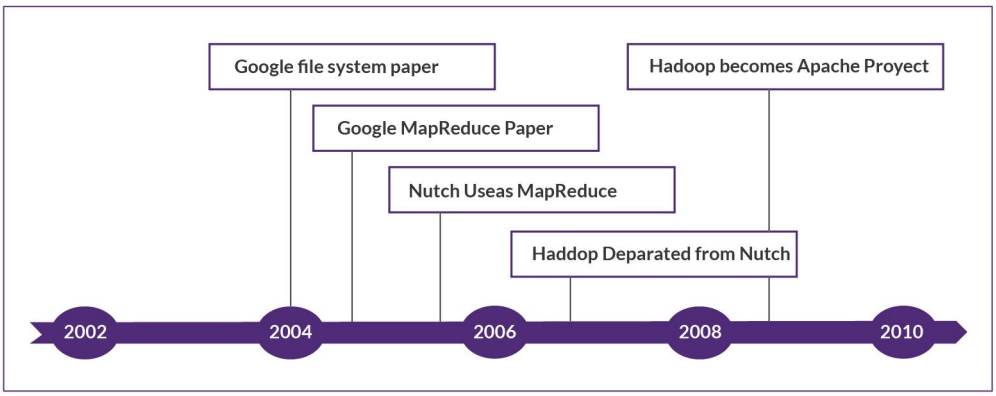


Figura 9. Linaje de Hadoop (Big Data in the Enterprise: Network Design Considerations).

Veamos esto con un simple ejemplo de una clásica situación de acceso y almacenamiento de la información para entender por qué con la tecnología tradicional no podríamos procesar el big data. Este simple ejemplo, muy sintetizado, permitirá poner luz sobre las ventajas de Hadoop.

EJEMPLO

Supongamos que disponemos de una computadora personal, la cual cuenta con un Disco Duro (disco rígido), el cual dispone una velocidad de acceso de 5.000Mbps (megabits/segundo) y que su capacidad de almacenamiento es de 1TB (1 Terabyte, 1.000.000 Megabytes). Si quisiéramos hacer una lectura completa del disco duro, la operación nos demoraría unos 200 segundos aproximadamente (el cálculo surge de hacer 1.000.000 MB / 5.000 = 200 segundos).

Supongamos un caso hipotético donde ahora la misma información la tenemos en 100 discos duros (a razón de 10GB/Disco) conectados de forma paralela; la lectura de 1TB de información llevaría un tiempo de 2(dos) segundos (10.000/5.000) es decir, hubiésemos reducido drásticamente la velocidad de acceso a la información.

La conclusión de este experimento es que, si tenemos más discos trabajando en forma paralela y la información está distribuida entre ellos (sin importar la posición), lograríamos reducir la velocidad de acceso (que originalmente era de 200 segundos). Y, si además, pensamos en algún nuevo componente o sistema que permita tener redundancias a fallas y que no permita la pérdida de información, tendremos todas las características fundamentales del conocido sistema Hadoop para almacenamiento y acceso a los datos, de manera masiva y en paralelo. De esto se trata Hadoop. Esta es su principal característica y su principal distinción.

Hadoop es un marco de trabajo que permite procesar grandes volúmenes de datos haciendo uso eficiente de los recursos (por eso es que muchos autores hablarán de reducción de costos) y que, además, incluye una serie de componentes para procesar de manera óptima el almacenamiento que se encuentra distribuido, permitiendo trabajar con datos estructurados y no estructurados y, además, a gran escala, es decir, varios millones de Terabytes (Petabytes, o incluso Zettabytes).

Otra característica importante de Hadoop, además del procesamiento distribuido, es que permite hacer implementaciones en equipamientos de hardware (servidores o incluso computadoras personales), de bajo costo, es decir que no se requieren grandes inversiones de dinero para implementar Hadoop.

Los sistemas operativos más utilizados por Hadoop son Linux y Windows, aunque existen otros como BSD y OS X.

“Hadoop se diseñó para correr en un gran número de máquinas que no comparten memoria ni discos. Eso significa que se pueden comprar un gran número de servidores, unirlos en un armario (rack) y ejecutar el software Hadoop en cada uno de ellos. Cuando se desea cargar todos los datos de su organización en Hadoop, lo que hace el software, es agrupar estos datos en piezas que se despliegan a continuación a través de sus diferentes servidores.

Hadoop sigue la traza donde residen los datos, haciendo múltiples copias de modo tal que los datos alojados en un servidor pueden ser replicados automáticamente a partir de una copia conocida.

En un sistema centralizado de base de datos, se tiene una gran unidad de disco conectada a cuatro, ocho, dieciséis o más procesadores, y este sistema aguanta mientras se tenga potencia suficiente. En un sistema clúster de Hadoop, cada uno de esos servidores tiene dos, cuatro, ocho o dieciséis CPU.

Puede ejecutar su trabajo de indexación, enviando su código a cada una de las docenas de servidores de su clúster, y cada servidor opera sobre su propia pequeña pieza de datos. Los resultados se entregan como si fuera un todo unificado” (Aguilar, 2013).

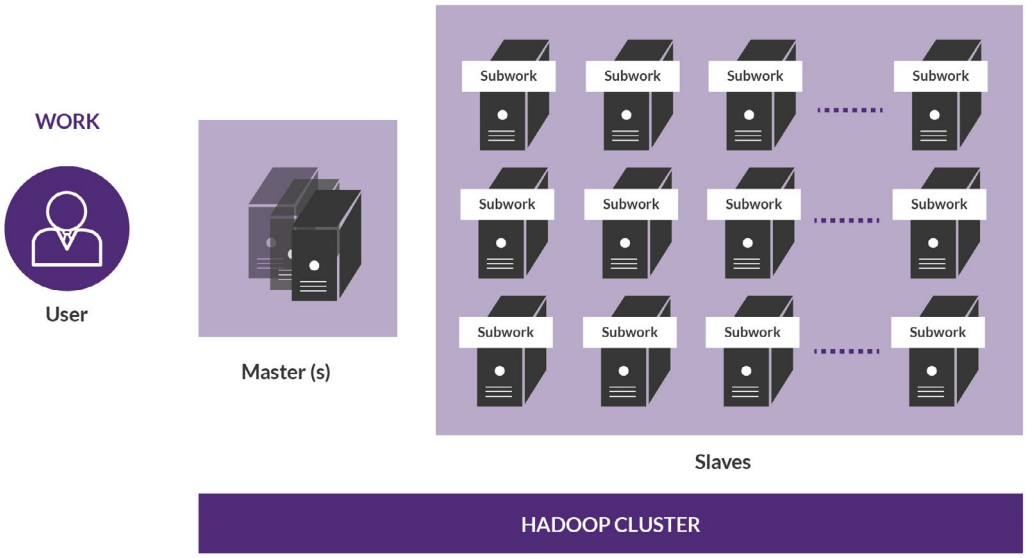


Figura 10. Hadoop Cluster (Apache Hadoop Introduction Tutorial - CloudDuggu).

Empresas como *Facebook*, *Amazon*, *Twitter*, *Linkedin* (entre otras) eligen este tipo de tecnologías para operar día a día.

Hadoop es un sistema de archivos distribuido cuya tarea principal es resolver el problema de almacenar la información que supera la capacidad de una única máquina. Para solucionar este problema, un sistema de archivos distribuido gestionará y permitirá el almacenamiento de los datos en diferentes máquinas conectadas a través de una red de modo que se haga transparente al usuario la complejidad de su gestión.

Hadoop se considera normalmente que consta de dos partes fundamentales: un sistema de archivos, el HDFS (*Hadoop Distributed File System*), y el paradigma de programación MapReduce. Hadoop, al contrario que los sistemas tradicionales, está diseñado para explorar a través de grandes conjuntos de datos y producir sus resultados mediante un sistema de procesamiento distribuido por lotes (*batch*).

Con el correr de los años, surgió la versión 2.0 de Hadoop, donde se incorpora una tercera parte fundamental llamada Yarn (*Yet Another Resource Negotiator*) la cual resulta ser una pieza fundamental que permite soportar varios motores de ejecución incluyendo MapReduce, y que proporciona un planificador agnóstico a los trabajos que se encuentran en ejecución en el clúster. También se encarga de proporcionar los recursos computacionales necesarios para los trabajos como

memoria o CPU. En definitiva, se agregó Yarn para optimizar el funcionamiento de Hadoop en todo sentido.

Con el crecimiento de los servicios Cloud y la computación en la nube, Hadoop se volvió aún más potente, porque permite implementar su software en equipos virtuales, lo que conlleva a una gran cantidad de ventajas técnicas y operativas. Por ejemplo, sería factible aumentar los recursos como memoria, CPU, o disco de esas máquinas virtuales o reducirlos según las necesidades del negocio. Además logra independizar el almacenamiento de lo que es el procesamiento.

En los inicios, Hadoop funcionaba muy bien. Resolvía los problemas antes mencionados, pero con el correr del tiempo, y con la implementación de proyectos de Big data, comenzaron a salir a la luz una serie de cuestiones técnicas que Hadoop no lograba resolver de manera tan eficiente. Es así que comienzan a nacer nuevos proyectos Open Source, y comienzan a crearse una serie de herramientas que logran resolver de manera mucho más eficiente, lo que Hadoop no puede. Nos referimos con esto al denominado **Ecosistema de Hadoop.**

6.1. HADOOP Y SU ECOSISTEMA

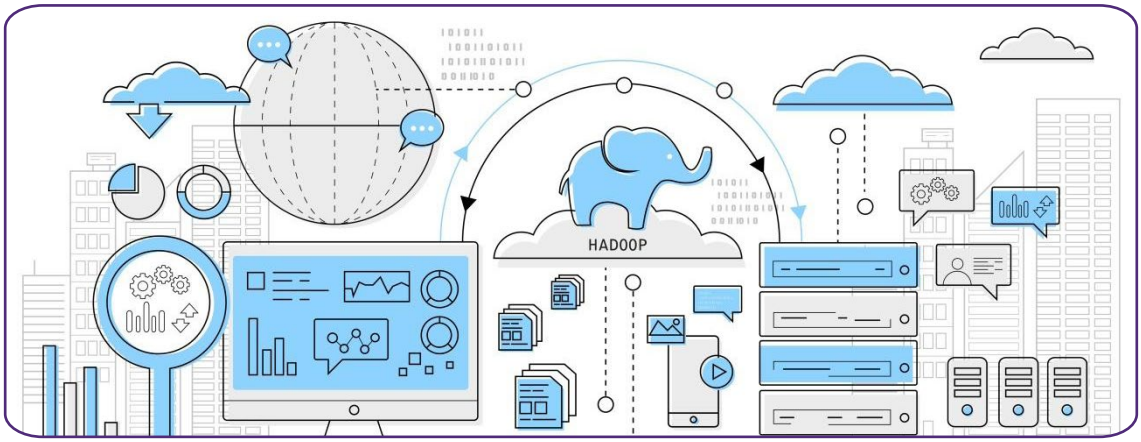


Figura 11. Ecosistema de Hadoop

NOTAS

Cuando antes hablábamos de Data Lakes... en verdad, debemos visualizar que esos data lakes están construidos con Hadoop.

Alrededor del framework de Hadoop, compuesto por las tres piezas fundamentales: **MapReduce, HDFS y Yarn**, han surgido todo un conjunto de tecnologías que la complementan y cumplen funciones específicas. Por ejemplo, existen tecnologías que facilitan la ingesta de datos hacia el clúster de Hadoop, otras que aceleran el procesamiento o bien facilitan la búsqueda de datos.

La característica que distingue al ecosistema de Hadoop son los componentes y herramientas que se incorporan con íconos de animales y nombres un tanto extraños. Veremos de qué se trata cada uno y pondremos luz sobre para qué sirven y dónde se utilizan.

Es importante mencionar que todo el ecosistema de Hadoop consiste en una serie de herramientas y aplicaciones que se pueden mostrar en capas. Típicamente, para comenzar la explicación se comienza con la explicación de la capa más baja del hardware, que es el almacenamiento, y luego se analizan el resto de las capas, con niveles de abstracción cada mas grandes, para dar lugar a las respuestas que los usuarios necesitan.

Sobre un hardware tradicional, un servidor o una computadora personal incluso, puede implementarse un sistema de archivos distribuidos que se denomina HDFS (*Hadoop Distributed File System*), el cual permite almacenar datos (tal como lo hemos explicado anteriormente) y eso se implementa sobre los NODOS del sistema y se ocupa de la gestión completa sobre cómo y donde almacenar los datos.

Por otra parte, y sobre la capa que acabamos de mencionar, se implementa lo que se denomina un gestor de recursos (YARN – *Yet Another Resource Negotiator*), del cual también hemos hablado anteriormente; quien en definitiva lo que hará es distribuir y planificar los trabajos en los diferentes nodos.

Por encima de este gestor Yarn, tendremos el modelo MapReduce que es quien se encarga de procesar los datos de una forma eficiente, y quien al mismo tiempo provee un modelo de programación.

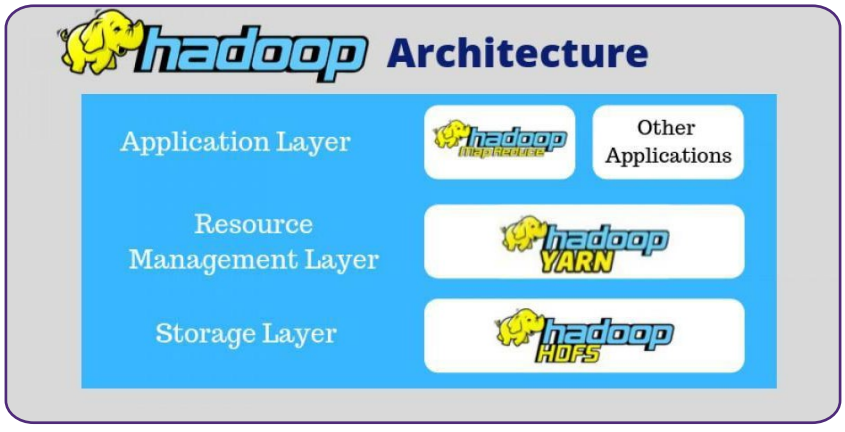


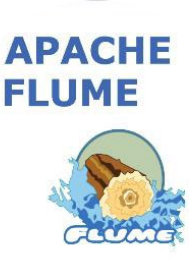
Figura 12. Arquitectura Hadoop (Hadoop Architecture: HDFS, Yarn & MapReduce).

Por encima de este entorno básico definido en la arquitectura hadoop, se introducen una serie de herramientas para facilitar el acceso y el procesamiento de todos esos datos.

Aquí es donde podemos introducir a **MongoDB o a Cassandra** como gestores de base de datos NoSQL (porque justamente nos permitirán gestionar cualquier tipo de dato, estructurados o no).

Estas bases de datos son herramientas que no están sobre el HDFS que mencionábamos antes, por lo que podrían funcionar perfectamente en cualquier equipo sin HDFS. En cambio, si utilizamos HDFS directamente y queremos montar un gestor de base de datos NoSQL sobre este HDFS podríamos utilizar Hbase como herramienta, ya que funcionaría igual que MongoDB y Cassandra, pero correo sobre HDFS y es una herramienta Hadoop base.

Hasta aquí lo que tenemos es nuestra arquitectura y ecosistema de Hadoop que nos permitirá almacenar la información.



Vale mencionar que para la realización de la ingesta de los datos nos encontraremos con herramientas tales como **Flume, Sqoop y Storm**.

Flume es un servicio que sirve para recopilar y mover grandes cantidades de datos entre los distintos nodos del sistema.

Sqoop básicamente es una herramienta que nos permite integrar datos que se encuentran en bases de datos tradicionales (relacionales) directamente a un sistema de base de datos de hadoop con HDFS como vimos anteriormente.

Storm es un sistema que se utiliza para resolver algunos problemas que inicialmente tenía Hadoop en relación con la ingesta de datos en tiempo real (streaming). Storm es un sistema muy eficiente que permite procesar mas de 1 millón de registros por segundo (por nodo).

En las capas superiores de la arquitectura Hadoop nos vamos a encontrar con aplicaciones y herramientas muy útiles tales como **Apache PIG y Apache HIVE:**

Apache Pig es una plataforma para el análisis de grandes conjuntos de datos que consta de un lenguaje de programación de alto nivel para expresar programas de análisis, junto con la infraestructura para la evaluación de los mismos. La característica sobresaliente de los programas de Pig es que su estructura es susceptible a la paralelización, lo que a su vez le permite manejar enormes cantidades de información. Mejora notablemente lo que se hace con MapReduce. Actualmente Pig es muy empleado en proyectos de big data.

Apache Hive es una tecnología distribuida diseñada y construida sobre Hadoop. Permite hacer consultas y analizar grandes cantidades de datos almacenados en HDFS, en la escala de petabytes. Tiene un lenguaje de consulta llamado HiveQL o HQL que internamente transforma las consultas SQL en trabajos MapReduce que ejecutan en Hadoop.

6.2. APACHE SPARK

Apache Spark es un framework de programación para procesamiento de datos distribuidos, el cual fue diseñado para ser rápido y de propósito general. Como su propio nombre indica, ha sido desarrollada en el marco del proyecto Apache, lo que garantiza su licencia Open Source.

Una de las grandes preguntas sobre Apache Spark es su relación con Hadoop. Muchos piensan que son competencia, cuando en realidad, Spark es la evolución natural de Hadoop, cuya funcionalidad es muy rígida y limitada en el sentido de que no aprovecha al máximo las capacidades del procesamiento distribuido.

Algunas de las evoluciones que supone Spark frente a su predecesor son el procesamiento en memoria que disminuye las operaciones de lectura/escritura, la posibilidad de análisis interactivo con SQL (similar a Hive en cierto modo) y la facilidad para interactuar con múltiples sistemas de almacenamiento persistente.
La ventaja principal y justamente la rapidez que caracteriza a Apache Spark es que al operar directamente sobre la memoria, agiliza notoriamente la performance respecto de Hadoop.

Es frecuente encontrar Spark en proyectos de big data Streming. Pero no debemos perder de vista que se ha vuelto un framework de referencia al igual que Hadoop, aunque spark es bastante más eficiente.

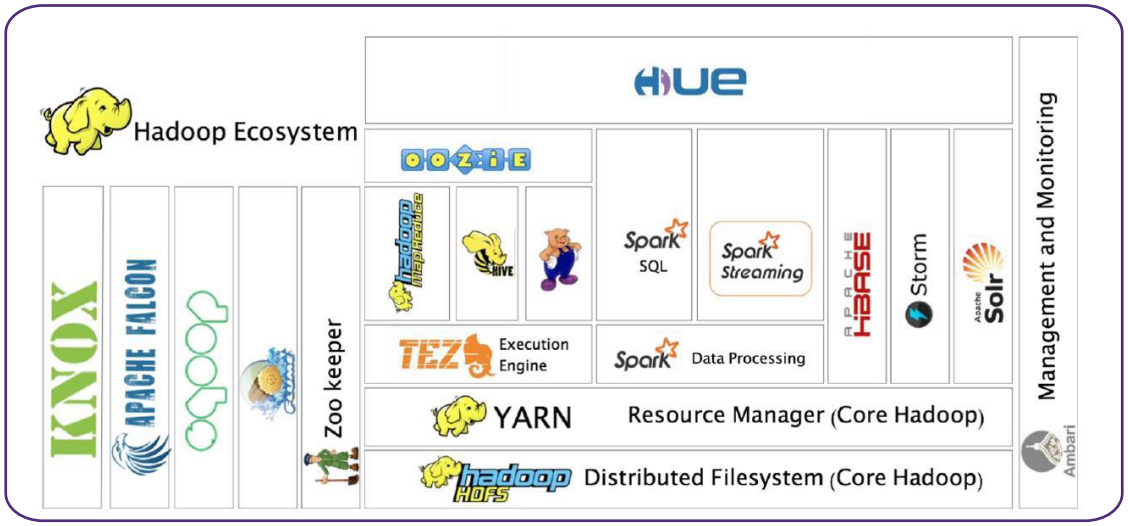


Figura 13. Síntesis del ecosistema de Hadoop (www.mikelnino.com).

BIBLIOGRAFÍA

Aguilar, Luis J. (2013). Big data, análisis de grandes volúmenes de datos en organizaciones. Ed. Alfaomega. Mexico.

Dean, J. and Ghemawat, S. (3 de octubre 2004) Mapreduce: Simplified Data Processing on Large Clusters.

GUIDE, A. Project management body of knowledge (pmbok® guide) 6th edition. Project Management Institute PMI.

Kenneth C. Laudon & Jane P. Laudon. Sistemas de Información Gerencial. 10ma edición. Pearson.

REFERENCIAS

Imágenes de portada: Shutterstock.