TEMA 71

CIENCIA DE DATOS. CICLO DE VIDA DE LOS PROCESOS DE MODELADO DE DATOS (ETL, PREPROCESADO, MODELADO, FUNDAMENTOS ESTADÍSTICOS, VALIDACIÓN, MLOPS). FUNDAMENTOS ESTADÍSTICOS. HERRAMIENTAS Y LENGUAJES. VISUALIZACIÓN DE DATOS.

Versión	30.1
Fecha de actualización	11/09/2024

ÍNDICE

1. Ciencia de datos

1.1. Introducción

La Ciencia de Datos tiene su origen en el año 1962, cuando el estadístico estadounidense **John W. Tukey**, conocido por el desarrollo de complejos algoritmos y el famoso diagrama de caja y bigotes (Box Plot), escribe y se cuestiona el futuro de la estadística como ciencia empírica. **Escribiría así en El futuro del análisis de datos (1962).**

Sin embargo, no sería hasta más adelante en **1974** cuando **Peter Naur**, científico danés conocido por sus trabajos en las ciencias computacionales y ganador del premio Turing en el año 2005, acuñara el término que actualmente conocemos.

Por su parte, **William S. Cleveland**, informático y estadístico estadounidense, famoso por sus aplica**cion**es en la visualización de datos, introdujo en **2001** a la Ciencia de Datos como una disciplina unificada y con independencia de lo que hasta ese momento se había conocido como Estadística.

La **ciencia de datos** es un campo académico interdisciplinario que utiliza estadísticas, computación científica, métodos, procesos, algoritmos y sistemas científicos para extraer o extrapolar conocimientos e ideas a partir de datos ruidosos, estructurados y no estructurados.

Por otro lado, la metodología **CRISP-DM (Cross-Industry Standard Process for Data Mining)** es un modelo de proceso estándar abierto que describe enfoques comunes utilizados por expertos en minería de datos. Es el modelo analítico más utilizado.

CRISP-DM fue concebido en 1996. En 1997 se puso en marcha como un proyecto de la Unión Europea bajo la iniciativa de financiación ESPRIT.

CRISP-DM divide el proceso de minería de datos en seis fases principales:

Comprensión del negocio

Comprensión de los datos

Preparación de los datos

Fase de Modelado

Evaluación

Implantación.

La ciencia de datos se utiliza para estudiar los datos de cuatro maneras principales:

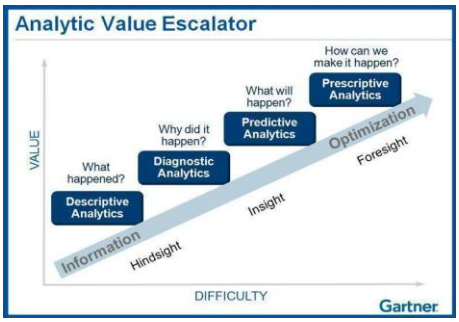
Análisis descriptivo: El análisis descriptivo examina los datos para obtener información sobre lo que ha ocurrido u ocurre en el entorno de datos. Se caracteriza por las visualizaciones de datos, como los gráficos

circulares, de barras o líneas, las tablas o las narraciones generadas.

Análisis de diagnóstico: El análisis de diagnóstico es un examen profundo y detallado de datos para entender por qué ha ocurrido algo. Se caracteriza por técnicas como el análisis detallado, el descubrimiento y la minería de datos o las correlaciones.

Análisis predictivo: El análisis predictivo utiliza los datos históricos para hacer predicciones precisas sobre los patrones de datos que pueden producirse en el futuro. Se caracteriza por técnicas como el machine learning, la previsión, la coincidencia de patrones y el modelado predictivo. En cada una de estas técnicas, se entrena a las computadoras para aplicar ingeniería inversa a las conexiones de causalidad en los datos. La ingeniería inversa utiliza todos los datos históricos y a partir de ellos genera patrones para obtener predicciones, frente al modelo estadístico que se basa en muestras, población e hipótesis para inferir relaciones entre las variables.

Análisis prescriptivo: El análisis prescriptivo lleva los datos predictivos al siguiente nivel. Predice lo que es probable que ocurra y sugiere una respuesta óptima. Usa análisis de gráficos, simulación, procesamiento de eventos complejos, redes neuronales y motores de recomendación de ML.



1.2. Tecnologías de la ciencia de datos

La ciencia de datos abarca el uso de tecnologías como:

Inteligencia artificial: los modelos de machine learning y el software relacionado con este se utilizan para los análisis predictivos y prescriptivos.

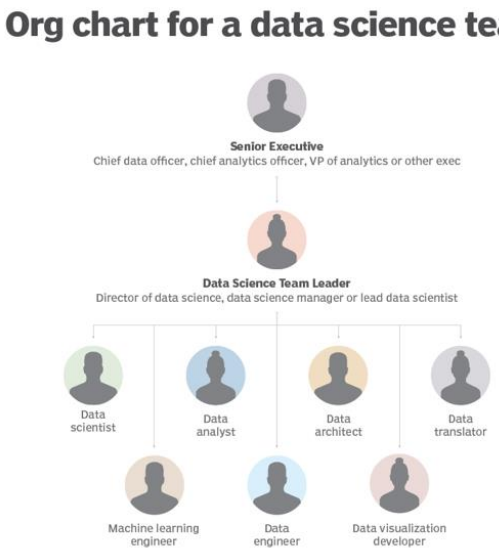
Computación en la nube y computación en el borde: las tecnologías en la nube ofrecen a los científicos de datos la flexibilidad y la capacidad de procesamiento necesarias para el análisis avanzado de datos.

Internet de las cosas: IoT se refiere a varios dispositivos que se pueden conectar de forma automática a Internet. Estos dispositivos recogen datos para iniciativas de la ciencia de datos. Generan datos masivos que se pueden utilizar para la minería y extracción de datos.

Computación cuántica: las computadoras cuánticas pueden realizar cálculos complejos a alta velocidad. Científicos de datos cualificados las utilizan para crear algoritmos cuantitativos complejos.

1.3. Roles

Un equipo de ciencia de datos puede contener una alta cantidad de perfiles, roles y responsabilidades. Algunos de las más importantes son:



- **Data Explorer**. Descubrir los datos del negocio.
- **Data Architect**. Organizan datos para analizar, plantear la estrategia de datos.
- **Data Steward**. Administrador de Datos. Vela por la calidad, seguridad y disponibilidad.
- **Data Engineer**. Almacenamiento y procesado de los datos.
- **Data Analyst**. Perfil previo a Data Scientists, abarcando una amplia gama de funciones que incluyen la minería, obtención y/o recuperación de datos, procesado, estudio y visualización.
- **Data Scientists**. Organizan los modelos de análisis y procesado. Manejan métodos estadísticos, algoritmos de aprendizaje automático y otras herramientas para analizar datos y crear modelos predictivos
- **Data Translator**. Híbrido entre negocio y datos, haciendo de puente entre las áreas más técnicas y las áreas más específicas de negocio.
- **Business Analyst**. Unión entre negocio e IT. Recopila e interpreta información de negocio.
- **Campaign Experts**. Transforman modelos a resultados.
- **Data Artist**. Científicos de datos que además son expertos en procesamiento gráfico y textual de flujos de datos.
- **Data Hygienists**. Limpian los datos y adaptan al formato del proyecto.
- **Chief Data Officer (CDO)**. Responsable de todos los equipos de Big Data.
- **Data Visualization Developer**. Crea informes operativos y analíticos, integrando múltiples conjuntos. Visualización datos, presentando hallazgos de forma lógica y fácil de interpretar.
- **Machine Learning Engineer**. También llamado Ingeniero de IA, colabora con científicos de datos para crear, implementar y mantener los algoritmos y modelos necesarios para el aprendizaje automático y las iniciativas de IA.
- **MLOps Engineer**. Se encarga de llevar a cabo un conjunto de prácticas que automatizan y simplifican los flujos de trabajo y los despliegues de machine learning (ML).

2. Ciclo de vida de los procesos de modelado de datos

Dentro del área de ciencia de datos, el número de procesos y procedimientos puede ser variable. Incluso algunas etapas pueden englobar y relacionar otras.

El ciclo de vida de los procesos de modelado de datos contiene:

Obtención de los datos: Los datos masivos suelen venir de múltiples fuentes, pueden ser de volúmenes diversos, se generan rápidamente y se debe comprobar que sean correctos.

Preprocesamiento de los datos: Se realiza un tratamiento inicial de los datos, donde se limpian y filtran aquellos datos que no cumplen criterios de calidad, no son de interés para el estudio, contienen errores…

Transformación e integración: Homogenizar los datos que provienen de múltiples fuentes para que sean comparables entre ellos. Esto puede deberse a la estructuración (datos en formato de tabla como APIs REST en formato json o xml) o a la no estructuración de los datos (datos en cualquier otro formato como texto, imágenes …).

Análisis de los datos: Procesar los datos usando diferentes algoritmos y métodos estadísticos para obtener resultados que respondan a las preguntas planteadas por los científicos de datos.

Interpretación de los datos: Es en este punto donde el científico de datos evalúa el resultado del análisis y aplica la experiencia que tiene en el campo para entender, completar y corregir la información.

Ingeniería de características. Implica la extracción y transformación de variables a partir de datos sin procesar.

Validación de los datos: Ver si estos datos son robustos o cambian por sesgos propios de los datos. Puede validarse de múltiples maneras: mediante datos externos al proceso, empleando técnicas diferentes a las empleadas en el estudio… pero siempre han de obtener un resultado similar a los obtenidos inicialmente para afirmar que los resultados son reales y no debido al azar o sesgos.

Diseñar nuevos análisis o experimentos en caso necesario: En el procedimiento científico esta parte es la que se define como “Validar la hipótesis”. En caso de que los datos no hayan sido validados o se necesite más información para poder obtener resultados concluyentes a las preguntas planteadas por los científicos de

datos, se incluyen mayor número de datos en los análisis o se reformulan los algoritmos para realizar otras preguntas a los datos.

Visualizar y presentar gráficamente los resultados de los datos: Es un proceso fundamental en cualquier trabajo con grandes bases de datos, el graficar de forma completa y con la mayor cantidad de capas posibles la información resultante. Las gráficas son formas rápidas de interpretar los datos para tomar decisiones y la tendencia en todos los artículos científicos y en la vida cotidiana en general, es a ir complicando y completando la cantidad de información que se ha obtenido en una única imagen.

2.1. ETL

La primera parte del proceso se denomina comúnmente ETL (Extract, Transform and Load). Este proceso permite mover datos provenientes de diferentes fuentes, además de formatearlos y limpiarlos, para finalmente guardarlos en otra base de datos, Datamart, Datalake o Datawarehouse.

Extracción de los datos de las fuentes de origen. Éstos pueden provenir de diferentes orígenes como BBDD, plataformas tecnológicas o protocolos de comunicaciones. Es más, estas fuentes tampoco suelen utilizar la misma organización ni formato.

Transformación de los datos, en la que se realizan diferentes pasos como la limpieza, donde los datos se pueden depurar, corregir o estandarizar.

Carga de los datos, que una vez depurados y "limpios", se incorporan (ej. estructura OLAP), ya sea sobrescribiendo la información antigua con nuevos datos o incorporando a los nuevos.

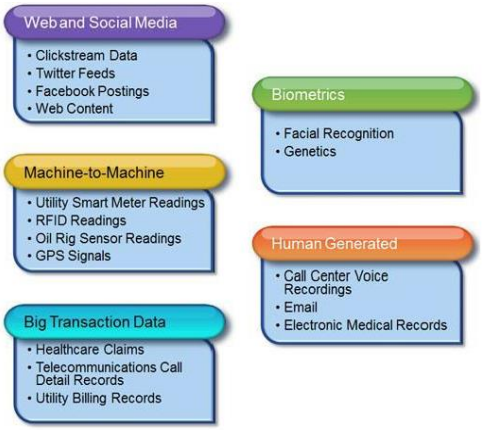
Algunas de las herramientas más utilizadas para llevar a cabo procesos ETL son:

Comerciales: Power Center, Oracle Data Integrator (ODI), SAS Data Integrator, Dastage, FME (datos geográficos), Azure Data Factory, SSIS, AWS Glue...

Open Source: Pentaho, CLover ETL, KETL, Talend Open Studio Integration

Como parte del proceso de ETL, la técnica de **Web Scraping** consistente en extraer información de sitios web a partir de técnicas de software. Su funcionamiento está vinculado a dos procesos: **Rastreo Web (crawling)**, también conocido como indexación y **Raspado (Scraping)** o extracción.

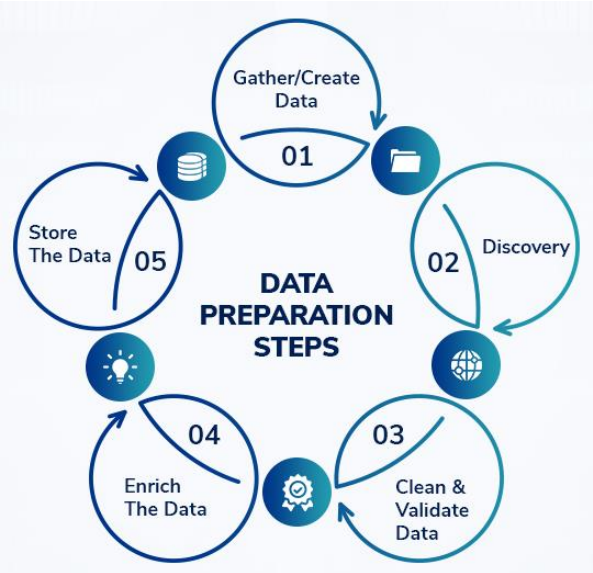
Algunas de las herramientas más utilizadas de Web Scraping son **Octoparse, ParseHub, Visual Scraper, 80legs, WebHarvy, Helium Screaper, HTTrack**, **librerías de Python (Beautiful Soup, Selenium, Scrapy**) entre otros.



2.2. Preprocesado

La preparación de datos, también conocida como "preprocesamiento", es el acto de limpiar y consolidar los

datos sin procesar antes de utilizarlos para realizar un análisis de negocio.



Recopilar/Crear datos: Es necesario un set de datos disponible.

Descubrimiento: Proceso de descubrimiento, buscando los conjuntos de datos importantes para el objetivo.

Limpiar y validar datos: También conocido como **Data Scrubbing o Data Cleansing**. Con los conjuntos de datos delineados, comienza la limpieza de los datos. Esto implicará completar los valores faltantes, eliminar información incorrecta y convertir los datos a un formato estándar…

Reducción de datos: Selección de datos relevantes. Algunas de las técnicas usadas **son la selección de características, selección de instancias, Data squashing (creación de datasets más pequeños a partir de uno más grande) o discretización**.

Enriquecer los datos: Los datos se agregan y conectan dentro de un conjunto, enriqueciéndolo y dándole una mejor comprensión de lo que significa para el negocio.

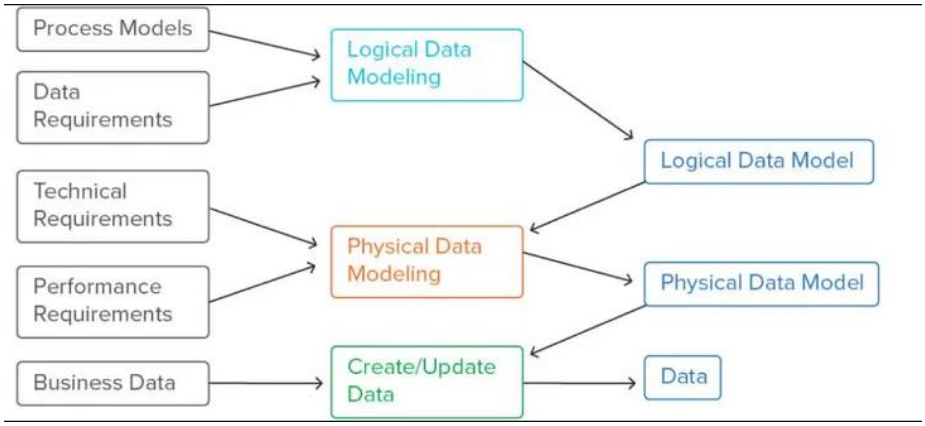
Almacenar los datos: Una vez preparados, los datos se almacenarán en un servidor hasta que llegue el momento de utilizarlos.

Algunas de las principales herramientas para realizar un preprocesamiento de datos son: **Librería Pandas y Numpy de Python, Weka, RapidMiner, R Studio, Azure Machine Learning, IBM Watson Studio.**

2.3. Modelado

El modelado de datos viene con el objetivo de producir datos de mayor calidad, estructurados y consistentes para ejecutar aplicaciones empresariales y lograr resultados consistentes. El modelado de datos en la ciencia de datos se puede denominar como un mecanismo diseñado para definir y ordenar datos para el uso y análisis por parte de ciertos procesos de negocio.

Uno de los objetivos del modelado en la ciencia de datos es crear el método más eficiente para almacenar información y, al mismo tiempo, proporcionar acceso completo e informes.



Existen tres tipos de modelo de datos:

Conceptual. Aplica un orden teórico sobre los datos según su existencia en relación con las entidades que se describen, a menudo conceptos o artefactos del mundo real.

Lógico. Un modelo de datos lógico suele ser el siguiente paso después del modelado conceptual de datos. Inherentemente, el proceso de modelado lógico intenta imponer el orden estableciendo valores clave, entidades discretas y relaciones en una estructura lógica, tomando la estructura semántica construida en la etapa conceptual.

Físico. Este es el paso de modelado de datos que divide los datos en las tablas, índices y clústeres reales necesarios para el almacenamiento de los datos. Este paso implica detenerse en más detalles con las claves primarias, las claves externas, las teclas de columna y las restricciones.

Existen distintos algoritmos de predicción o técnicas y métodos de modelado de ciencia de datos que se pueden emplear para realizar el análisis:

Árboles de Decisión. Son modelos de clasificación muy utilizados que tratan de encontrar la variable que permite dividir el dataset en grupos lógicos que son más diferentes entre sí. Cada árbol se va descomponiendo en distintas ramas y hojas que representan cada clasificación en función de las condiciones que se van seleccionando hasta llegar a la resolución del problema.

Redes neuronales. La Inteligencia Artificial y el Deep Learning han puesto muy de moda esta técnica tan sofisticada de reconocimiento de patrones que imita las neuronas del cerebro humano ya que es capaz de modelar relaciones extremadamente complejas.

Máquinas de Vectores de Soporte (SVM). Son algoritmos de aprendizaje automático supervisado de cara a reconocer patrones, estando relacionados con problemas de clasificación o regresión.

Análisis Bayesiano. Se trata de una inferencia estadística en la que las evidencias u observaciones se emplean para actualizar o inferir la probabilidad de que una hipótesis pueda ser cierta.

Regresión Logística. Las regresiones logísticas son utilizadas para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictivas.

Regresión Lineal. La regresión lineal consiste en utilizar la línea recta que muestre el "mejor encaje" de todos los puntos de los valores numéricos (datos) disponibles, para predecir nuevos valores. El modelo vendrá dado por la recta que minimice en mayor medida la distancia al conjunto de dichos puntos. También se llama el método de los mínimos cuadrados por ser la distancia de mínimos cuadrados la más empleada para obtener la recta.

Series Temporales y Data Mining. Combina una mezcla de técnicas de data mining tradicional como sampleado, clustering y árboles de decisión, con otras de forecasting con el fin de mejorar las predicciones

sobre datos recopilados como ventas por meses o trimestres, llamadas por día, o visitas a nuestra web por hora.

K-Vecinos más Cercanos. Algoritmo de agrupamiento o clustering. Consiste en reconocer patrones para conocer la probabilidad de que un elemento pertenezca a una clase según su cercanía en el espacio a los elementos de esa clasificación.

Ensemble Models. Algoritmos de boosting y bagging. Crea un nuevo modelo entrenando varios modelos similares combinando los resultados para mejorar la precisión, reducir la varianza y los sesgos e identificar el mejor modelo para usar con nuevos datos.

Potenciación del Gradiente. Lleva a cabo un resampling (método de remuestreo) de nuestro dataset para generar unos resultados que formen una media ponderada del conjunto de datos.

Modelos de Respuesta Incremental. Suele utilizarse para reducir el Churn (rotación de clientes) o comprobar la efectividad de diferentes acciones de Marketing. Se modela el cambio de probabilidad causado por una acción.

2.4. Validación

Los métodos de validación, también conocidos como **resampling**, son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso de los datos de entrenamiento.

Algunas de las estrategias de validación son:

Validación Simple. Consiste en repartir aleatoriamente las observaciones disponibles en dos grupos, uno se emplea para entrenar al modelo y otro para evaluarlo.

Leave One Out Cross-Validation (LOOCV). Es un método iterativo que se inicia empleando como conjunto de entrenamiento todas las observaciones disponibles excepto una, que se excluye para emplearla como validación.

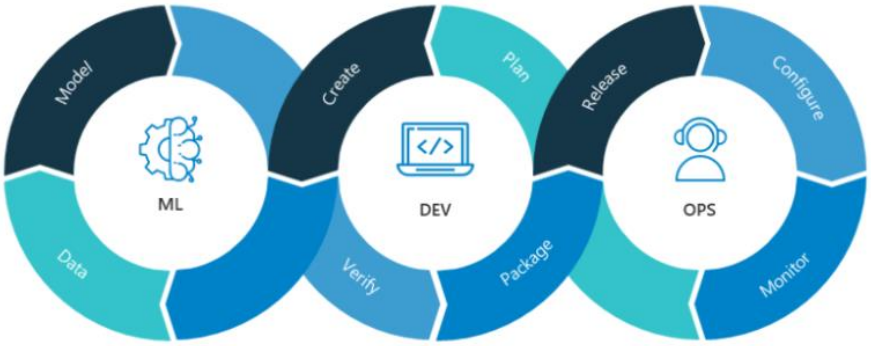
K-Fold Cross-Validation. El método K-Fold Cross-Validation es también un proceso iterativo. Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración.

Repeated k-Fold-Cross-Validation. Similar al método k-Fold-Cross-Validation pero repitiendo el proceso completo n veces.

Bootstrapping. Una muestra bootstrap es una muestra obtenida a partir de la muestra original por muestreo aleatorio con reposición, y del mismo tamaño que la muestra original. Muestreo aleatorio con reposición (resampling with replacement) significa que, después de que una observación sea extraída, se vuelve a poner a disposición para las siguientes extracciones.

2.5. MLOps

Machine Learning Operations (MLOps) es una de las nuevas tendencias en ciencia de datos, una parte de la cultura DevOps. Por ello, está estrechamente relacionada con los modelos de aprendizaje automático (machine learning) y la automatización de procesos.



MLOps (Machine Learning Operations)

El objetivo de MLOps es desarrollar, entrenar y desplegar modelos en producción de una manera fiable de ML con procedimientos automatizados que integren a los equipos de Data, desarrolladores, seguridad e infraestructura. La puesta en producción de los modelos es desplegada de forma conjunta por los Científicos de Datos y los ingenieros DevOps y Aprendizaje Automático.

MLOps (Machine Learning Operations) y DevOps (Development Operations) son dos términos relacionados, pero distintos. Ambos se centran en mejorar la eficiencia, la calidad y la escalabilidad en el desarrollo de software, pero difieren en los objetivos específicos y los métodos utilizados.

Dentro de MLOps existen tres niveles de implementación de Machine Learning:

Data: datos, fase, ingestión, curado, etc.

Model: testing, evaluación de los modelos, empaquetado y como se van a despegar.

Code: el código, donde se ejecuta todo el modelo en sí.

La tarea principal que resuelven las operaciones de aprendizaje automático es la gestión de datos a gran escala. Gracias a este modelo, se pueden resolver cuestiones como:

La falta de profesionales especializados en datos.

Mejora de comunicación entre los científicos de datos y los equipos técnicos.

Facilita y acelera la gestión de datos.

Gestiona los ciclos de vida de una empresa

Genera flujos de trabajo automáticos que mejoran las operaciones internas

Reduce el sesgo en la toma de decisiones, ya que iguala la representación de los diferentes sectores entre los datos

Algunas de las herramientas más usadas son **Mlflow, DVC, Kubeflow, ONNX, TensorFlow Extended, Apache Airflow, Cookiecutter, Metaflow, Jupyter, PyTorch o Psycaffold.**

AIOps (Inteligencia Artificial en operaciones de TI) es la aplicación de algoritmos de aprendizaje automático para generar procesos automatizados que ayudan a los equipos de TI a mejorar el flujo de trabajo y la colaboración. Entre otras cosas, este modelo logra automatizar los diagnósticos de incidentes y busca el causante con ML. Filtra los datos y reduce los errores.

3. Fundamentos Estadísticos

Uno de los objetivos de la ciencia de datos y machine learning es crear modelos predictivos a partir de un histórico. Es por ello que la estadística juega un papel clave.

La estadística se divide en dos partes diferenciadas como son la **estadística descriptiva y la inferencia estadística.**

La estadística descriptiva es el conjunto de métodos estadísticos que describen y/o caracterizan un grupo de datos. La estadística inferencial busca deducir y sacar conclusiones acerca de situaciones generales más allá del conjunto de datos obtenidos.

	Estadística descriptiva	Estadística inferencial
Definición	Métodos empleados para resumir las características clave de datos conocidos.	Métodos que implican el uso de datos muestrales para hacer generalizaciones o inferencias acerca de una población.
Objetivos	Caracterizar un grupo de datos Examinar tendencias o distribuciones	Examinar diferencias entre grupos. Examinar si las variables están asociadas. Comparar promedios entre grupos. Predecir una variable a partir de otra.
Métodos de análisis	Medidas de tendencia central: • Media • Mediana • Moda Medidas de variabilidad: • Varianza • Desviación estándar • Rango • Frecuencia	• t-test • Análisis de varianza • Correlación • Regresión
Áreas de aplicación	Ciencias naturales y sociales	Ciencias sociales y naturales
Ejemplos	• Características de pacientes que son atendidos en un hospital. • Media y distribución de la edad, peso y altura de los estudiantes de un colegio.	• Predecir la aparición de demencia en personas según su estado cardiovascular. • Probar que un medicamento sirve para el tratamiento de una enfermedad.

3.1. La estadística descriptiva

La **estadística descriptiva o deductiva** es la técnica matemática que obtiene, organiza, presenta y describe un conjunto de datos con el propósito de facilitar el uso, generalmente con el apoyo de tablas, medidas numéricas o gráficas.

Elementos de la estadística descriptiva. Fuente: *Ramos et. al. (p. 16)*

Algunos de los fundamentos estadísticos o conceptos son:

Dispersión. En el marco de una variable determinada, se produce una distancia entre valores. A esta diferencia se le llama dispersión.

Promedio. El promedio es la media y por lo tanto la tendencia dentro de una variable. Es decir, el resultado de la división de la suma de valores entre el número de ellos.

Sesgo. El sesgo o curtosis de una variable es la cualidad de la curva de datos. Es decir, el valor que tiene que

ver con la distancia y proximidad de los datos con respecto al promedio.

Gráficos. La materialización o presentación de los datos desprendidos de un análisis se realiza en forma de representación gráfica. Existe variada tipología de gráficos: en barras, círculo, lineales, polígonos...

Asimetría. Los diferentes datos de una variable se distribuyen con respecto al promedio de una forma determinada; este valor se llama asimetría.

3.2. Estadística inferencial

También llamada estadística inductiva. **La estadística inferencial** es la parte de la estadística que comprende métodos y procedimientos que por medio de la inducción determina propiedades de una población estadística, a partir de una parte de esta. Es decir, observa una muestra de datos y extrae conclusiones que aplica al conjunto a través de inferencias. Este tipo de enfoque, al ser resultado de un cálculo de tipo probabilístico, conlleva cierto margen de error.

Elementos de la estadística inferencial. Fuente: *Ramos et. al. (p. 17)*

Se puede categorizar a la estadística inferencial en dos grandes grupos:

Pruebas de hipótesis. Se trata de validar aquellas conclusiones que se han construido respecto a esa porción de datos estudiados.

Intervalos de confianza. Se llama intervalo de confianza a un intervalo dentro del cual se estima, con un determinado nivel de confianza, que estará el valor de cierto parámetro poblacional desconocido. Formalmente, los extremos del intervalo, se calculan a partir de los datos de una muestra correspondiente a un estimador del parámetro poblacional.

La estadística inferencial se subdivide a su vez en dos grandes tipos: estadística **paramétrica** y **no paramétrica**.

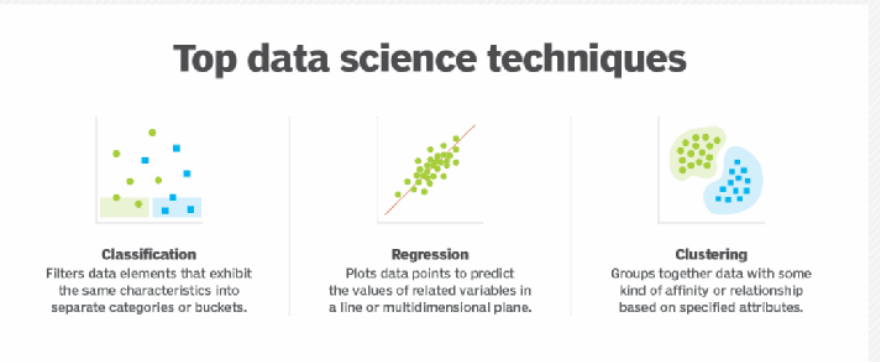
4. Herramientas y Lenguajes

Algunas de las principales técnicas usadas en la ciencia de datos son:

Clasificación, que separa los elementos de un conjunto de datos en diferentes categorías (algortimos supervisados).

Regresión, traza los valores óptimos de las variables de datos relacionadas en una línea o un plano.

Agrupación, que agrupa puntos de datos con una afinidad o atributos compartidos (algoritmos no supervisados).



Los lenguajes de programación para ciencia de datos permiten a los desarrolladores escribir programas que se ocupen de análisis de datos masivos, como estadísticas y máquina de aprendizaje.

La siguiente lista no pretende ser un conjunto exhaustivo de lenguajes y herramientas para el desarrollo de ciencia de datos, dado que hoy en día existe una colección extremadamente extensa, sino un conjunto de lenguajes y herramientas de referencia inicial.

Python. Python es un lenguaje de programación de alto nivel, interpretado y de propósito general, muy adecuado para el desarrollo rápido de aplicaciones. Tiene una sintaxis simple y fácil de aprender que permite una curva de aprendizaje pronunciada y reducciones en los costos de mantenimiento del programa. Hay muchas razones por las que es el idioma preferido para la ciencia de datos. Por mencionar algunos: potencial de secuencias de comandos, verbosidad, portabilidad y rendimiento.

Algunas de las librerías más utilizadas son:

Python Numpy para almacenar y operar. Numpy nos permite crear todo tipo de estructuras numéricas, múltiples dimensiones, permite transformarlas, operar aritméticamente, filtrar y es útil muchas veces para la inicialización de datos aleatorios. Otra de las funcionalidades es la de dar soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas.

Python Pandas para manipular los datos. Con ella podemos leer archivos o bases de datos de múltiples fuentes (csv, sqlite, sqlserver, html, json, xml) y hacer operaciones entre las columnas, ordenar, agrupar, dividir, pivotar, totalizar. Nos ayuda a detectar valores nulos, detectar ouliers, duplicados y también hacer merge ó joins entre distintos orígenes. También nos permite guardar fácilmente nuestro nuevo dataset o dataframe.

Plotly. Plotly (Plot.ly) es una herramienta de análisis y visualización de datos. Plotly proporciona herramientas gráficas, analíticas y estadísticas en línea, así como bibliotecas de gráficos científicos para Python, R, MATLAB , Perl, Julia, Arduino y REST. Plotly está construido con Python y el framework Django, con un front-end que utiliza JavaScript y la biblioteca de visualización D3.js, HTML y CSS.

Scikit-learn cubre las necesidades al momento de preprocesar datos, hacer transformaciones y crear modelos de ML. Incluye varios algoritmos de clasificación, regresión y análisis de grupos entre los cuales están máquinas de vectores de soporte, bosques aleatorios, Gradient boosting, K-means y DBSCAN. Está diseñada para interoperar con las bibliotecas numéricas y científicas NumPy y SciPy.

Matplotlib. Biblioteca para la generación de gráficos en dos dimensiones, a partir de datos contenidos en listas o arrays en el lenguaje de programación Python.

Entre los lenguajes y plataformas más utilizados en la ciencia de datos se encuentran:

R. Se utiliza principalmente para procesamiento de datos estadísticos y gráficos. Aunque no está destinado a desarrollar aplicaciones completas, como sería el caso de Python, R se ha vuelto muy popular en los últimos años debido a su potencial para la minería de datos y el análisis de datos. R es capaz de realizar todo tipo de trabajo de procesamiento de datos, incluido el modelado lineal / no lineal, clasificación, pruebas estadísticas, etc.

Spyder. IDE genérico está destinado principalmente a científicos y analistas que también necesitan codificar. Para hacerlos más cómodos, no se limita a la funcionalidad IDE, sino que también proporciona herramientas para la exploración / visualización de datos y la ejecución interactiva, como se puede encontrar en un paquete científico

PyCharm. Tiene un editor de código inteligente con búsqueda inteligente, finalización de código y detección y corrección de errores

Julia. Julia es un lenguaje de uso general y se puede usar para escribir cualquier aplicación, muchas de sus características son adecuadas para el análisis numérico y la ciencia computacional.

Scala. Scala son las siglas de Scalable Language (lenguaje escalable) y es un lenguaje de programación multiparadigma de propósito general diseñado para integrar las características y funcionalidades de los lenguajes orientados a objetos y funciones. La compilación del código fuente de Scala produce bytecode Java para su ejecución en una JVM

Jupyter Notebook. Jupyter Notebook no es un lenguaje de programación en sí si no un intérprete. Es una aplicación web de código abierto que permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Es una herramienta que se utiliza para diversas actividades, como la limpieza y transformación de datos, la simulación numérica, la modelización estadística, la visualización de datos y el aprendizaje automático, entre otras. A nivel de cuaderno, la herramienta es compatible con más de 40 lenguajes de programación, incluidos Python, R, Julia y Scala.

Trifacta. Trifacta es una plataforma que permite explorar y preparar los datos para su análisis. Trifacta funciona con plataformas de datos locales y en la nube, y está diseñado para permitir a los analistas explorar, transformar y enriquecer datos crudos y heterogéneos en formatos limpios y estructurados para el análisis de autoservicio.

5. Visualización de datos

La visualización de datos es la representación gráfica de la información. Mediante el uso de elementos visuales, como gráficos, mapas e infografías, las técnicas de visualización permiten comunicar de una manera accesible la información a los diferentes usuarios.

Los datos nos cuentan historias (**Storytelling**), y debemos narrarlas adecuadamente dependiendo de quiénes serán los consumidores de la información.

- **Los dashboards** o paneles de datos consisten en la consolidación de la representación gráfica de todos aquellos indicadores claves de éxito (**KPI**) de una organización, y está orientado a la toma de decisiones estratégicas de una empresa.
- **KGI** (Key Goal Indicator) o "Indicador de objetivo clave", es una métrica que establece para medir si se ha logrado un objetivo comercial importante a largo plazo.
- **KPI** (Key Performance Indicator) o "Indicador clave de rendimiento", señala cuán efectivos son sus métodos actuales para acercarlo a su objetivo final.
- **OKR** (Objective and Key Result) significa "Objetivo y resultado clave" y es una metodología colaborativa para establecer y rastrear objetivos.
- **KSF (Key Success Factor)** o factor clave de éxito. KSF se refiere a los factores críticos que determinan si una empresa tiene éxito en el logro de sus objetivos.
- Algunas de las técnicas usadas para la visualización de los datos son:
- **Gráficos lineales**. Esta es una de las técnicas más básicas y comunes utilizadas. Los gráficos de líneas muestran cómo las variables pueden cambiar con el tiempo.

- **Gráficos de área**. Este método de visualización es una variación de un gráfico de líneas; muestra múltiples valores en una serie de tiempo, o una secuencia de datos recopilados en puntos consecutivos e igualmente espaciados en el tiempo.
- **Gráfico de dispersión**. Esta técnica de visualización utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos. Un diagrama de dispersión puede sugerir varios tipos de correlaciones entre las variables estudiadas con un intervalo de confianza determinado.
- **Mapas de árbol**. Este método muestra datos jerárquicos en un formato anidado. El tamaño de los rectángulos utilizados para cada categoría es proporcional a su porcentaje del total. Los mapas de árbol se usan mejor cuando hay varias categorías presentes, y el objetivo es comparar diferentes partes de un todo.
- **Pirámides de población**. Esta técnica utiliza un gráfico de barras apiladas para mostrar la compleja narrativa social de una población. Se utiliza mejor cuando se intenta mostrar la distribución de una población.

Las herramientas más relevantes para la visualización de datos son: Kibana, SpagoBI, Grafana, Tableau, IBM Watson Analytics, MS Power BI, Qlik, Microstrategy, Matplotlib (librería de python para proyectos de visualización), Data Science Notebooks, Zoho Analytics, Domo...

Ilustración 3. Cuadrante de Gartner de herramientas para visualización de datos (fuente: Gartner)

6. Minería de datos

La minería o explotación de datos es la etapa de análisis de "Knowledge Discovery in Databases" o KDD. Su objetivo es extraer información de un conjunto de datos que de algún modo se encuentra oculta, y transformarla en una estructura comprensible para su uso posterior. Además de la etapa de análisis en bruto, supone aspectos de gestión de datos y de bases de datos, de procesamiento de los mismos, del modelo y de las consideraciones de inferencia, de métricas de intereses, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea.

Data Mining Group ha estandarizado el lenguaje **PMML** (Predictive Model Markup Language).

Ejemplos de soluciones de Minería de Datos:

- Libre: Pentaho Data Mining (Weka), JHepWork, Orange Data Mining.
- Propietario: Cubos OLAP Excel, Enterprise Miner, Clementine, SAP Miner, Tableau y ThoughtSpot.

Según su nivel de abstracción, el conocimiento contenido en los datos puede clasificarse en distintas categorías y requerirá una técnica más o menos avanzada para su recuperación:

- **Conocimiento evidente**: Información fácilmente recuperable
  - Técnica Consulta SQL
- **Conocimiento multidimensional:** abstracción para considerar los datos con una cierta estructura
  - técnica: OLAP (análisis multidimensional).
  - se trata de reinterpretar una tabla con n atributos independientes como un espacio n-dimensional.
- **Conocimiento oculto**: Información no evidente, desconocida a priori y potencialmente útil
  - Técnica: minería de datos como reconocimiento de regularidades o algoritmos de aprendizaje automático.
- **Conocimiento profundo**: Información que resulta imposible de recuperar.

Para la elaboración de cuadros de mando y obtención de indicadores, es aconsejable aplicar la norma **UNE 66175:2003** Sistemas de gestión de la calidad. Guía para la implantación de sistemas de indicadores.

6.1. Características

- Trabaja con la información oculta.
- Suelen ser soluciones con una arquitectura cliente-servidor.
- Poseen gran variedad de herramientas para la extracción de la información.
- Es habitual hacer uso de un procesamiento paralelo que acelere el proceso debido a la existencia de una gran cantidad de datos.
- Produce cinco tipos de información:
  - Asociaciones
  - Secuencias
  - Clasificaciones
  - Agrupamientos
  - Pronósticos

6.2. Técnicas

Se basa en el uso de diferentes técnicas de aprendizaje automático:

- **Redes neuronales**: inspirado en la forma en que funciona el sistema nervioso de los animales. Cuando una red neuronal se compone de gran número de capas, se llama Deep Learning.
- **Árboles de decisión:** utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.
- **Reglas de inducción**: en base a cálculos estadísticos es posible generar reglas útiles que siguen un patrón de si - entonces y pueden aplicarse a los datos a analizar para extraer conclusiones y hacer pronósticos.

- **Sistemas expertos**: simulan capacidad de tomar decisiones de expertos humanos en solución de problemas. Tres tipos:
    - Basados en reglas previamente establecidas (if… then…).
    - Basados en casos o CBR (Case Based Reasoning).
    - Basados en redes bayesianas, modelo probabilístico de variables aleatorias.
- **Algoritmos genéticos**: se inspiran en los procesos de evolución natural y evolución genética. Cada solución candidata tiene un conjunto de propiedades (cromosomas o genotipos) que pueden ser mutados y alterados.
- **Clustering o agrupamiento:** se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes.

Estas técnicas se basan en el uso de algoritmos que se pueden clasificar en:

- **De verificación**: verifica la validez de la información que se le presenta.
- **Supervisados**: se clasifica en dos categorías (clasificación y regresión).
    - predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.
    - Algunos de los algoritmos supervisados más representativos son regresión, árboles de decisión, redes neuronales, estadísticos, etc.
- **No supervisados:**
    - se descubren patrones y tendencias en los datos.
    - Algunos de los algoritmos no supervisados más representativos son clustering, redes de asociación, detección de anomalías etc.

7. Big Data

Gartner define al Big Data como: “activos de información caracterizados por su volumen elevado, velocidad elevada y alta variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y la toma de decisiones en las organizaciones”, identificando las tres Vs que caracterizan este concepto:

- **Volumen**: se trabaja con gran cantidad de datos.
- **Velocidad**: los datos en movimiento como consecuencia de la creación de datos en tiempo real.
- **Variedad**: diferentes tipos de fuentes y de datos.

En la actualidad se han propuesto nuevas Vs al conjunto:

- **Veracidad** de los datos: En referencia a la incertidumbre de los datos, es decir, al grado de fiabilidad de la información.
- **Viabilidad**: Capacidad de una organización para utilizar de forma eficaz el gran volumen de datos que maneja.
- **Visualización de los datos**: La forma en que los datos son presentados una vez que se procesan.
- **Valor de los datos**: Se refiere al valor que se puede obtener de ellos cuando se transforman en información.

UIT-T Y.3600 - recomendación UIT que hace referencia a grandes volúmenes de datos - requisitos y capacidades basados en la computación en la nube.

8. Entornos Hadoop o similares

Inspirado en GFS (Google File System) y en el paradigma MapReduce.

Para comprender completamente cómo funciona Hadoop debemos entender sus tres componentes principales.

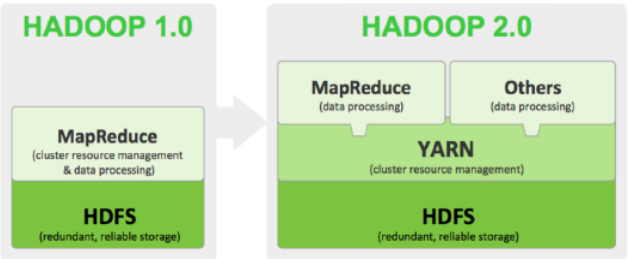
- **MapReduce**
- **HDFS** (Hadoop Distributed File System)
- **Yarn** (Yet Another Resource Negotiator)

Además de estos componentes, el módulo **Hadoop Common** aporta las utilidades y librerías comunes del proyecto que soportan el resto de componentes y permiten desarrollar las aplicaciones.

MapReduce es un paradigma de programación que permite una escalabilidad masiva en cientos o miles de servidores en un clúster de Hadoop. Como componente de procesamiento, MapReduce es el corazón de Apache Hadoop . El término "MapReduce" se refiere a dos tareas separadas y distintas que realizan los programas Hadoop: map y reduce. El primero es el trabajo de mapeo (map), que toma un conjunto de datos y lo convierte en otro conjunto de datos, donde los elementos individuales se dividen en tuplas (pares clave / valor). El trabajo de reducción (reduce) toma la salida de un mapa como entrada y combina esas tuplas de datos en un conjunto más pequeño de tuplas. Como implica la secuencia del nombre MapReduce, el trabajo de reducción siempre se realiza después del trabajo de mapeo.

HDFS (Hadoop Distributed File System) es el componente principal del ecosistema Hadoop. Esta pieza hace posible almacenar data sets masivos con tipos de datos estructurados, semi-estructurados y no estructurados. Proporciona la división de los datos en bloques que necesita MapReduce para ejecutar sus fases map y reduce. Está optimizado para almacenar grandes cantidades de datos y mantener varias copias en el clúster para garantizar una alta disponibilidad y tolerancia a fallos.

Yarn (Yet Another Resource Negotiator) es una pieza fundamental en el **ecosistema Hadoop 2.0**. En Hadoop 2.0, YARN toma las capacidades de gestión de los recursos que residían en MapReduce y las empaqueta para que puedan ser utilizados por los nuevos motores de procesado. Con YARN, se permite ejecutar varias aplicaciones en Hadoop, todos compartiendo una gestión común de los recursos. MapReduce se convierte ahora en una librería Hadoop es decir una aplicación que reside en Hadoop y deja la gestión de recursos del clúster para el componente YARN.



Hadoop Core está compuesto por: Common (utilidades) + HDFS (BD) + **MapReduce** (Procesado)

- **Hadoop Common**. JAR y Scripts que son las utilidades del sistema.
- **HDFS**. Hadoop Distributed File System.
  - Construido para Apache Nutchweb y ahora es subproyecto de Apache Hadoop.
  - En Java, sobre Linux normalmente. Para aplicaciones con grandes volúmenes de datos.
  - Diseñado para escalar a Petabytes. Tamaño de bloque superior al habitual de 64 MB.
  - No se comporta bien para accesos aleatorios a archivos, ni cuando éstos se actualizan frecuentemente.Tolerante a fallos por redundancia y diseñado para implementarse en HW de bajo coste.
  - Protocolo WebDAV sobre HTTP para exponer HDFS.

- Apache Casandra es más eficiente que HDFS, pero el modo batch que junta ficheros es óptimo en comparación.

- Hadoop **MapReduce**
  - Computación paralela en datos repartidos en varios nodos. Bibliotecas en C++, Java y Python.
  - "Map" es para entrada, realiza asignación clave – valor por cada dato introducido.
  - "Reduce" agrupa resultados que coincidan con la asignación clave – valor.
  - Map + Reduce transforman lista de pares (CLAVE, VALOR) en lista de valores.
  - Arquitectura maestro-esclavo
    - Nodos del clúster JRE 1.6 o superior. Por SSH.
    - Maestro administra nombres y regula acceso con el NameNode.
      - Compuesto por Jobtracker, Tasktracker, Namenode, Datanode.
      - EditLog. Es el registro del Namenode. Metadatos de los nodos slave.
      - Datanote se encuentra en la Slave y se encarga de almacenar los datos.
      - Rack es la combinación de nodos de datos. Máximo 40 nodos máster. Cada rack tiene switch.
      - Archivo se divide en bloques (64 o 128 MB).
      - Namenode recibe Heartbeat y BlockReport del DataNode.

8.1. Distribuciones Hadoop

- **Cloudera**. Los primeros y con propia consola de instalación.
- **HORTONWORKS**. Distribuye sin SW propietario adicional, culpables de innovaciones y YARN.
- **MAPR**. Reemplaza al HDFS y usa sistema propio MapRFS. Está integrado en Ubuntu. Gratuita pero faltan algunas características.

8.2. Ecosistema Hadoop

CAPTURA Y MANIPULACIÓN DATOS

- LUCENE
  - Librería en Java para buscar textos. Indexar cualquier texto.
  - Funciona en texto plano, hay plugins que permiten la indexación y búsqueda en Word, PDF, XML o HTML.
- SQOOP. Sql-to-Hadoop
  - Para bulk data entre Hadoop y sistemas de almacenamiento con datos estructurados, como BD relacionales.
  - Permite importar tablas individuales o BD de SQL a HDFS o Hive.
  - Genera clases Java para interactuar con los datos importados.
- CHUKWA
  - Recolecta, analiza y visualiza logs. Herramientas para visualizar y monitorizar los datos capturados.

- Agentes + Colectores (Sink files) + Job MapReduce (.done los procesados) + HICC (Care Center – Portal).

- FLUME
  - En Java y distribuida para capturar y mover grandes volúmenes de diferentes orígenes a repositorio, como logs.
  - Flume y Chukwa son proyectos parecidos. Chukwa está pensado para ser usado en batch.
  - Nodos agente (instalados en servers, gestión flujo datos) y nodos colectores (agrupan flujo de datos).
- KAFKA
  - Plataforma distribuida de transmisión de datos para publicar, almacenar, procesar flujos y suscribirse en tiempo real.
  - LinkedIn para manejar 1.4 billones de mensajes por día. Ahora es Open source.
  - Diseñada para manejar flujos de varias fuentes y distribuirlos a usuarios.
  - Basado en P2P. Distribuido: Se divide en varios nodos para su ejecución dentro de un clúster.
  - Escalado horizontal: Se puede escalar en muchos nodos en relativo poco tiempo.
  - Patrón publicación-suscripción, masivamente escalable concebida como registro de transacciones distribuidas.
  - Cuándo utilizar Apache Kafka:
    - Arquitecturas de streaming de datos en tiempo real. Streams, agregación de registros, ingesta datos en Spark.
    - Envío, intercambio, transformación de datos entre aplicaciones.
    - Recopilar grandes datos (logs, changelogs, actividad web, monitoreo de métricas, seguimiento de clicks).
    - Funcionamiento como caché persistente.
    - Proporciona durabilidad en microservicios in-memory.
    - Proporciona eventos a CEP (Complex Event Streaming System) y a sistemas IoT/IFTTT.
    - Fundamentales para alta productividad y escalabilidad. Puede reducir la latencia a milisegundos.
    - IoT, Monitorización IT, RRSS o comercio electrónico. Los datos aumentan exponencialmente y pueden sobrecargar.

ALMACENAMIENTO

- HIVE.
  - DWH para agregación y análisis de datasets almacenados en Hadoop.
  - Suite BD que permite Resumen, Consultas ad-hoc y análisis. Permite ETL
  - Acceso a BD con HDFS o Apache HBASE.
  - Consultas con MapReduce y con Lenguaje HiveQL.
  - Tiene interfaces JDBC/ODBC para integrarse con herramientas de BI.
  - No está diseñado para cargas OLTP, consultas en tiempo real o actualizaciones a nivel de fila.
  - Hive es mejor para análisis de datos tradicionales. Pig Latin es mejor para un pipeline o investigar datos en bruto.
- HBASE
  - BD orientada a columnas de Hadoop. Funciona sobre HDFS. Distribuido, no permite SQL.
  - Para escrituras/lecturas en tiempo real y acceso aleatorio en grandes volúmenes.

- IMPALA

  - Base de datos SQL sobre Hadoop

- Proporciona la capacidad de realizar consultas concurrentes y de baja latencia para analítica y BI.

TRATAMIENTO DATOS

- OOZIE
  - Librería para definir secuencias de Jobs (diagrama de flujo). Compatible con XML.
  - Motor de workflow (Java Map-Reduce, Streaming Map-Reduce, Pig, Hive, Sqooq…) con reglas.
- APACHE PIG
  - De Yahoo y adaptado por Apache.
  - Análisis con PigLatin (SQL Like). Traduce consultas del usuario a Map-Reduce.
  - Entorno fácil de programación convirtiendo las paralelizaciones en dataflows.
- JAQL
  - De IBM, es un lenguaje de consulta para JSON y texto plano. Funcional y declarativo.
  - Select, join, group y filtrar datos almacenados en HDFS.
- STORM
  - MapReduce y Hadoop no son en tiempo real, Storm sí.
  - Código libre y abierto. Twitter lo usa.
  - En vez de aplicar un nodo MapReduce, se aplican Topologías master-slave. Grafos de computación.
- GIRAPH. Procesamiento sobre grafos. Usado en Facebook. Procesa 1 billón de aristas en 4 min.
- APACHE MAHOUT
  - Implementación de algoritmos de Machine Learning en máquinas distribuidas con HDFS.
  - Algoritmos: Filtrado colaborativo, Clustering, Clasificación, Frequent pattern mining…
- APACHE SPARK.
  - Framework de computación en clúster open-source en Scala y del ecosistema Hadoop. El más usado en Analítica de datos.
  - ARQUITECTURA KAPPA: Kafka + Spark + NoSQL + Scala.
  - Se gestiona desde la interfaz web y combina diferentes modelos de datos en la misma aplicación.
  - Paralelismo de Datos implícito, tolerancia a fallos y orientado a velocidad.
  - Aparición por problemas en MapReduce. Velocidad por ir en memoria y tener un DAG que soporta flows acíclicos. Permite en disco para grandes.
  - 100 veces más de velocidad en memoria y 10 veces en disco con MapReduce.
  - MapReduce no es óptimo para Real Time processing, Filtros y Joints por el Key value pattern, OLTP por las altas latencias, gráficos (Hadoop usa Giraf como librería) o iteraciones como Kmeans.
  - Spark es perfecto para lo que no llega MapReduce, se puede usar sobre Hadoop Map-Reduce para batch.
  - Muchos lenguajes y librerías.
  - Componentes: Spark Core, Spark SQL, Spark Streaming y Spark MLib.
  - RDD es una encapsulación de un dataset muy grande. En Spark se trabaja a nivel de RDD. Spark distribuye automáticamente los RDD y los paraleliza.

ADMINISTRACIÓN

- AMBARI.
  - Interfaz de usuario en web para gestión de clústeres Hadoop.

    - Incluye: HadoopHDFS, MapReduce, Colmena, HCatalog, HBASE, Zookeeper, Oozie, Pig y Sqoop.
- ZOOKEEPER
    - Infraestructura centralizada para gestión de configuraciones, naming, sincronización, grupos de servicios…
    - Interfaz simple. Elige automáticamente los nodos líderes.
    - Cola de mensajes entre nodos. Se garantiza la comunicación.
- YARN
    - Gestor de recursos. Decide la asignación de tareas en función de los recursos disponibles en los nodos.
    - Planificación flexible y gestión de recursos sobre el HDFS.
    - Yahoo utiliza YARN para planificar jobs en 40.000 servidores.
    - JobTracker: Gestión recursos (Resource Manager) y Planificación o Monitorización (Application Master).
    - Convierte MapReduce en una librería, no el lenguaje principal.
- TEZ. Aplica en YARN una perspectiva de grafos
- AVRO
    - Sistema de serialización de datos para procesarlos y almacenarlos; mejora de rendimiento.
    - La serialización puede ser en texto en plano, JSON o en binario.
    - Optimizado para minimizar el espacio en disco necesario para nuestros datos.

9. Bases de datos NOSQL

Frente a las bases de datos relacionales aparecen las bases de datos NoSQL que se caracterizan por:

- No utilizar SQL como lenguaje de consultas. Algunos tipos utilizan JSON, CQL (Cassandra), GQL (Big Table)…
- No utilizan estructuras fijas como tablas para el almacenamiento.
- No permiten operaciones Join. La solución es desnormalizar datos o usar JOIN a nivel aplicación.
- Emplean arquitecturas distribuidas con Master-Slave.

Las principales ventajas derivadas del empleo de este tipo de bases de datos son:

- Se ejecutan en máquinas con pocos recursos y por tanto de menor coste.
- Alta escalabilidad horizontal: para mejorar el rendimiento sólo hace falta añadir más nodos.
- Manejan gran cantidad de datos.
- No generan cuellos de botella. (SQL tiene que transcribir cada sentencia).

Razones para usarlas:

- El volumen de los datos crece muy rápidamente en momentos puntuales.
- La escalabilidad de la solución relacional no es viable.
- Elevados picos de uso del sistema por parte de los usuarios en múltiples ocasiones.

10. Tipos de bases de datos NoSQL

Se distinguen principalmente los siguientes tipos de BBDD NoSQL:

- **Clave-valor:** cada elemento se asocia con una clave única. Ej: Redis, DynamoDB, Aerospike, etc.
- **Orientadas a Documentos:** Datos semi estructurados o documentos en XML, JSON o BSON. Las más versátiles. Ej: Mongo DB, Couch DB, Raven DB, Apache Jena, Pivotal GemFire o MarkLogic Server.
- **Orientadas a Grafos:** Nodos y aristas para representar los grafos. Muy bien para modelos con muchas relaciones como redes o conexiones sociales. Ej: Neo4j, Infinite Graph, Orient DB, Arango DB, Virtuoso, DEX/Sparksee, AllegroGraph, OrientDB, SonesGraphDB, InfoGrid, HyperGraphDB.
- **Orientadas a objetos:** la información se almacena como objetos. Ej: Versant, Object DB, etc
- **Orientadas a columnas**: estos sistemas almacenan los datos cambiando el enfoque transaccional de filas a columnas. Ej: Cassandra, Hbase, etc.
- BD NoSQL Search. Índices Invertidos y documentos basados en los proyectos Lucene y Solr. Similar a las de documentos pero por índices. Ejemplos: Elasticsearch, MeiliSearch.

Algunas de las BBDD NOSQL más extendidas: **REDIS, MongoDB, Apache Cassandra, ELASTICSEARCH**.