

Aplicación de técnicas de Data Mining para la optimización de arquitecturas cloud



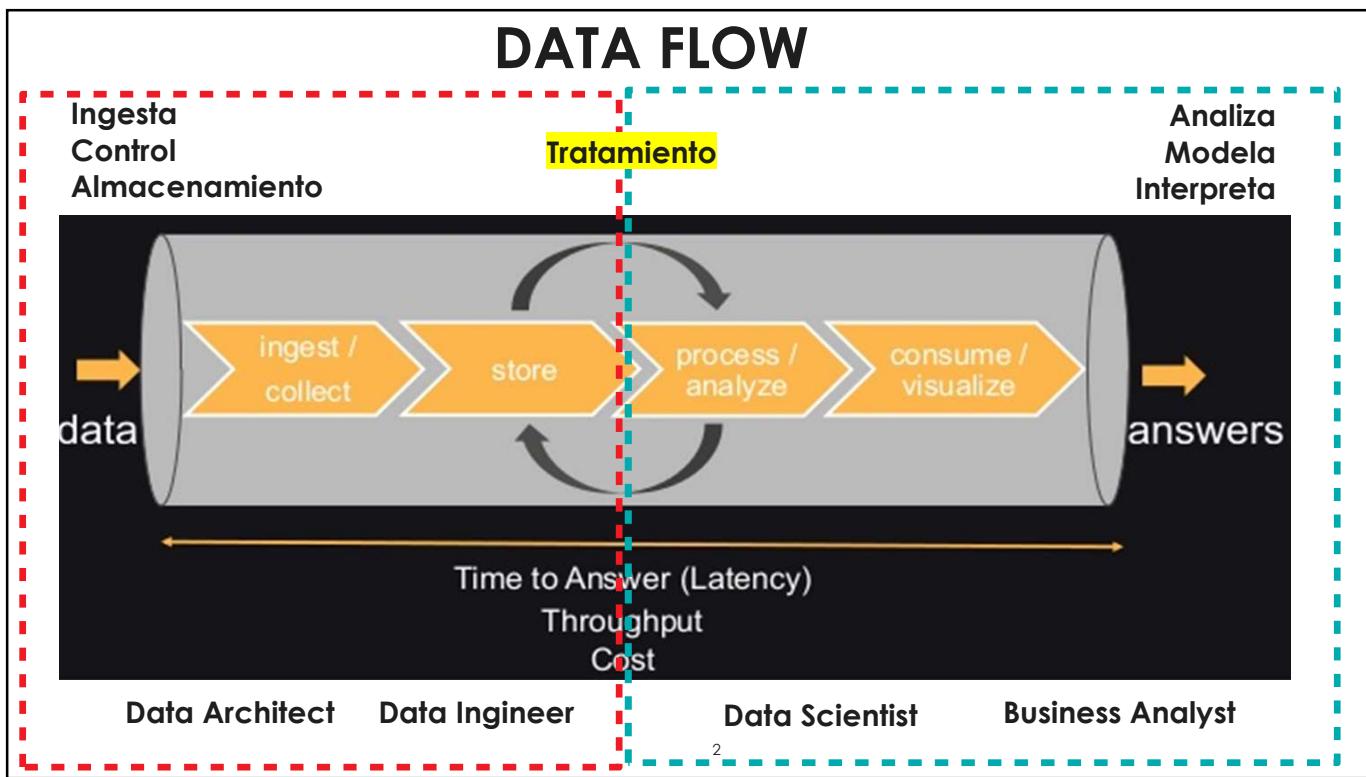
Autor: José María Escalante Fernández

- **Unidad 1 – ETL vs. ELT**
- **Unidad 2 – Relación entre Arquitecto, Ingeniero y Analista de datos**
- **Unidad 3 – Data Mining Techniques**
- **Unidad 4 – Modelo de procesos orientados a l análisis de datos**
- **Unidad 5 – Analítica de LOGs**

Unidad 1 – ETL vs. ELT

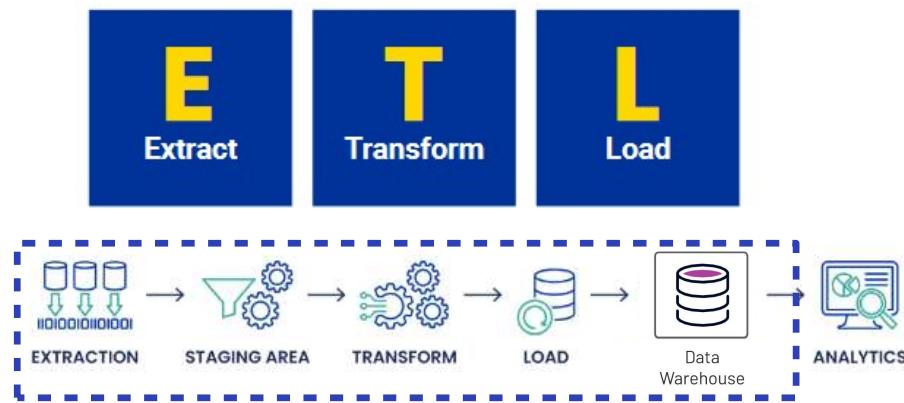


1



2

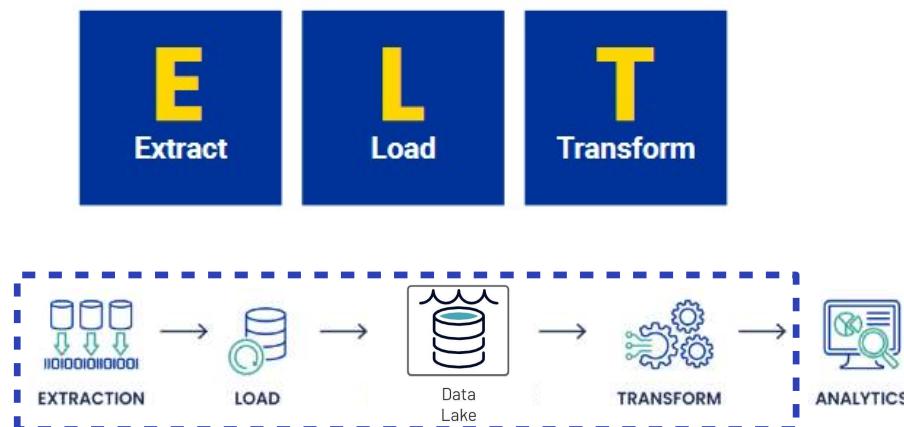
PROCESO ETL



El proceso de transformación de datos es llevado antes del almacenamiento de datos.

3

PROCESO ELT



El proceso de transformación de datos después del almacenamiento de datos.

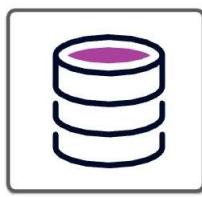
4

2

¿Cuál es la diferencia entre un Data Warehouse y un Data Lake?

5

¿Cuál es la diferencia entre un Data Warehouse y un Data Lake?



Data
Warehouse

Almacén de datos **estructurados y filtrados**, procesados **para un propósito concreto**.



Data
Lake

Almacén de datos **en bruto**, que todavía **no tiene una finalidad definida**.

6

ETL: pros and cons

PROS

- Ciclo de integración de datos extendido.
- Reducción el riesgo de exposición de datos sensible.
- Proceso clásico entre los analistas.
- Necesita menos almacenamiento de datos.

CONS

- Esquema de almacenamiento poco flexible (SQL).
- Requiere transformaciones complejas.
- Perdida de información por procesado inadecuado.
- Proceso más lento.

7

7

ELT: pros and cons

PROS

- Velocidad de almacenamiento.
- Procesamiento parcial de datos en paralelo.
- Posibilidad de reprocesar los datos.
- Escalabilidad.
- Menos mantenimiento.

CONS

- Herramientas de análisis más costosas.
- Combinación de bases de datos mixtas (SQL y NoSQL).
- Exposición de datos.
- Necesidad de gran almacenamiento de datos.

8

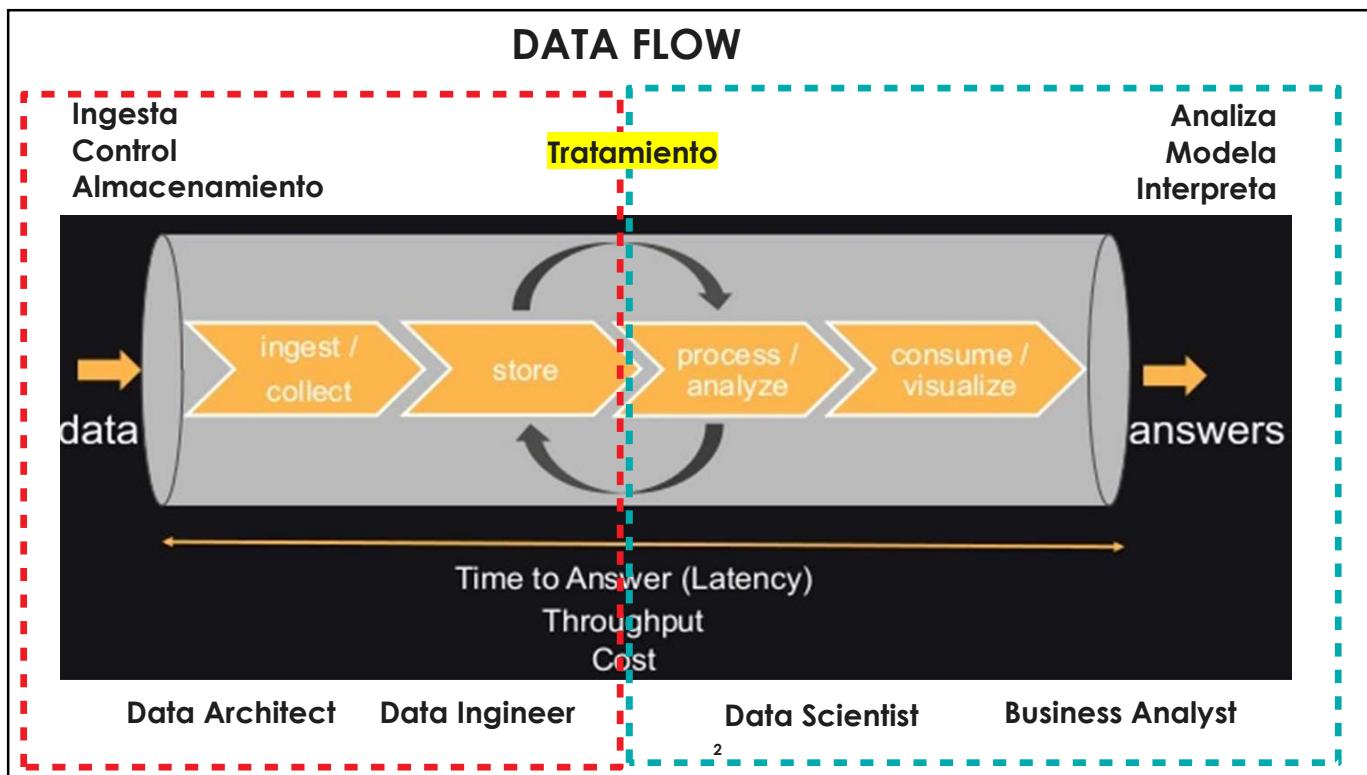
8

4

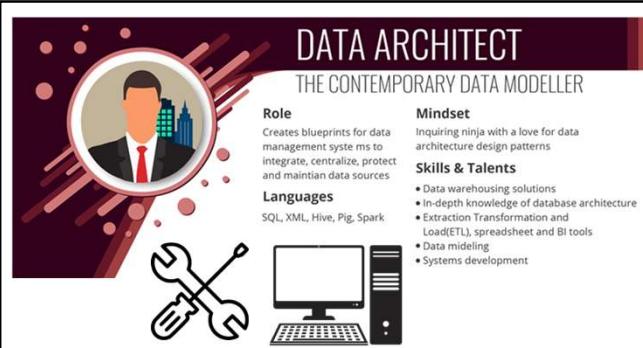
Unidad 2 – Relación entre Arquitecto, Ingeniero y Analista de datos



1



2



DATA ARCHITECT
THE CONTEMPORARY DATA MODELLER

Role Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources	Mindset Inquiring ninja with a love for data architecture design patterns
Languages SQL, XML, Hive, Pig, Spark	Skills & Talents
	<ul style="list-style-type: none"> • Data warehousing solutions • In-depth knowledge of database architecture • Extraction Transformation and Load(ETL), spreadsheet and BI tools • Data modeling • Systems development

Proporciona el marco fundamental para organizar y estructurar los datos dentro de una organización.

Se centra en el diseño de sistemas y estructuras que permiten la captura, almacenamiento y gestión eficiente de los datos.

Las principales actividades del arquitecto de datos son:

- **Modelos de Datos:** Definen la estructura y las relaciones entre los diferentes tipos de datos dentro de la organización, como los modelos relacionales, dimensionales o de grafos.
- **Almacenamiento de Datos:** Determina dónde y cómo se almacenarán los datos, ya sea en bases de datos relacionales, data lakes, almacenes de datos en la nube u otras soluciones de almacenamiento.
- **Integración de Datos:** Se refiere a los procesos y tecnologías utilizados para integrar datos de diversas fuentes y sistemas, garantizando la coherencia y la calidad de los datos.
- **Gobernanza de Datos:** Establece políticas y procedimientos para garantizar la seguridad, privacidad, calidad y cumplimiento normativo de los datos.

3



DATA ENGINEER
SOFTWARE ENGINEERS BY TRADE

Role Develops, constructs, tests and maintains architectures (such as databases and large scale processing systems)	Mindset All-purpose everyman
Languages SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl	Skills & Talents
	<ul style="list-style-type: none"> • Database systems (SQL & NO SQL based) • Data modeling & ETL tools • Data APIs • Data warehousing solutions

Implementan y gestionan los sistemas y procesos necesarios para el flujo eficiente de datos .

Responsable de construir y mantener las infraestructuras de datos implicando la captura, almacenamiento, procesamiento y distribución de grandes volúmenes de datos.

Sus principales responsabilidades incluyen:

- **Extracción, Transformación y Carga (ETL):** Desarrollar y mantener pipelines de datos para extraer datos de diversas fuentes, transformarlos en un formato adecuado y cargarlos en el sistema de almacenamiento.
- **Desarrollo de Arquitecturas de Datos:** Diseñar e implementar arquitecturas de datos escalables y tolerantes a fallos, utilizando tecnologías como Hadoop, Spark, Kafka, entre otras.
- **Optimización de Rendimiento:** Identificar y resolver cuellos de botella en el flujo de datos, optimizando el rendimiento de los sistemas de almacenamiento y procesamiento.
- **Automatización y Monitorización:** Desarrollar herramientas y procesos automatizados para gestionar y monitorizar pipelines de datos, garantizando la fiabilidad y disponibilidad del sistema.

4

DATA SCIENTIST
AS RARE AS UNICORNS

Role Cleans, massages and organizes (big) data	Mindset Curious data wizard
Languages R, SAS, Python, Matlab, SQL, Hive, Pig, Spark	Skills & Talents
	<ul style="list-style-type: none"> • Distributed computing • Predictive modeling • Story-telling and visualizing • Math, Stats, Machine Learning

A small icon of a person working at a laptop is also present.

Se enfoca en la extracción de conocimientos significativos y acciones prácticas a partir de los datos.

Utilizan técnicas estadísticas, algoritmos de machine learning y análisis de datos para descubrir patrones, predecir tendencias y tomar decisiones informadas.

Sus principales actividades incluyen:

- **Análisis Exploratorio de Datos:** Explorar y visualizar datos para identificar patrones, tendencias y relaciones ocultas que puedan ser relevantes para el negocio.
- **Modelado Predictivo:** Construir modelos predictivos utilizando técnicas de machine learning para predecir eventos futuros o clasificar datos en categorías específicas.
- **Optimización y Experimentación:** Desarrollar y ejecutar experimentos para optimizar procesos y tomar decisiones basadas en datos.
- **Comunicación de Resultados:** Presentar hallazgos y recomendaciones de manera clara y efectiva a partes interesadas no técnicas, utilizando visualizaciones y narrativas convincentes.

5

Una relación muy especial

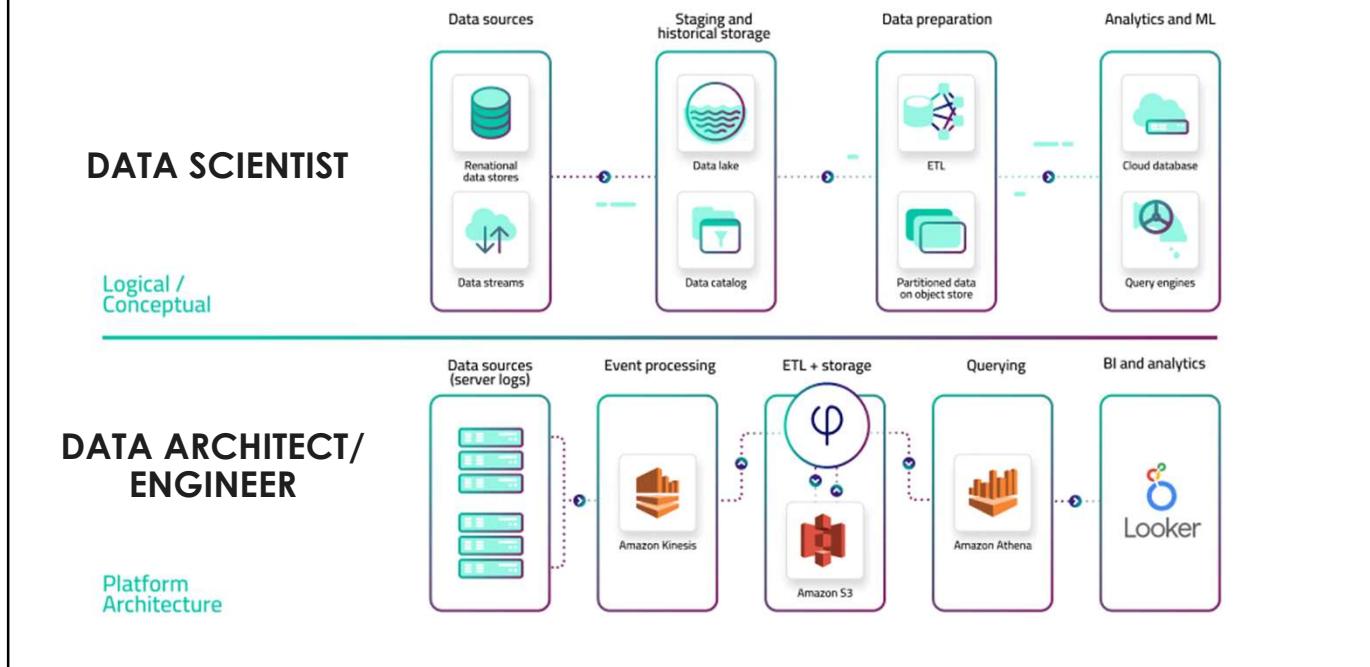
Aunque el Data Architect, Data Engineer y Data Scientist tratan disciplinas distintas, están estrechamente interrelacionadas y se complementan entre sí en el proceso de gestión y aprovechamiento de los datos.

- **Data Architect** proporciona el marco conceptual y estructural para organizar y gestionar los datos
- **Data Engineer** se encarga de implementar los sistemas y procesos necesarios para la ingestión y control de datos atendiendo al marco conceptual creado por el Data Architect.
- **Data Scientist** utiliza la infraestructura y mecanismo de almacenamiento e ingestión de datos para el análisis y extraer conocimientos significativos y generar valor.

6

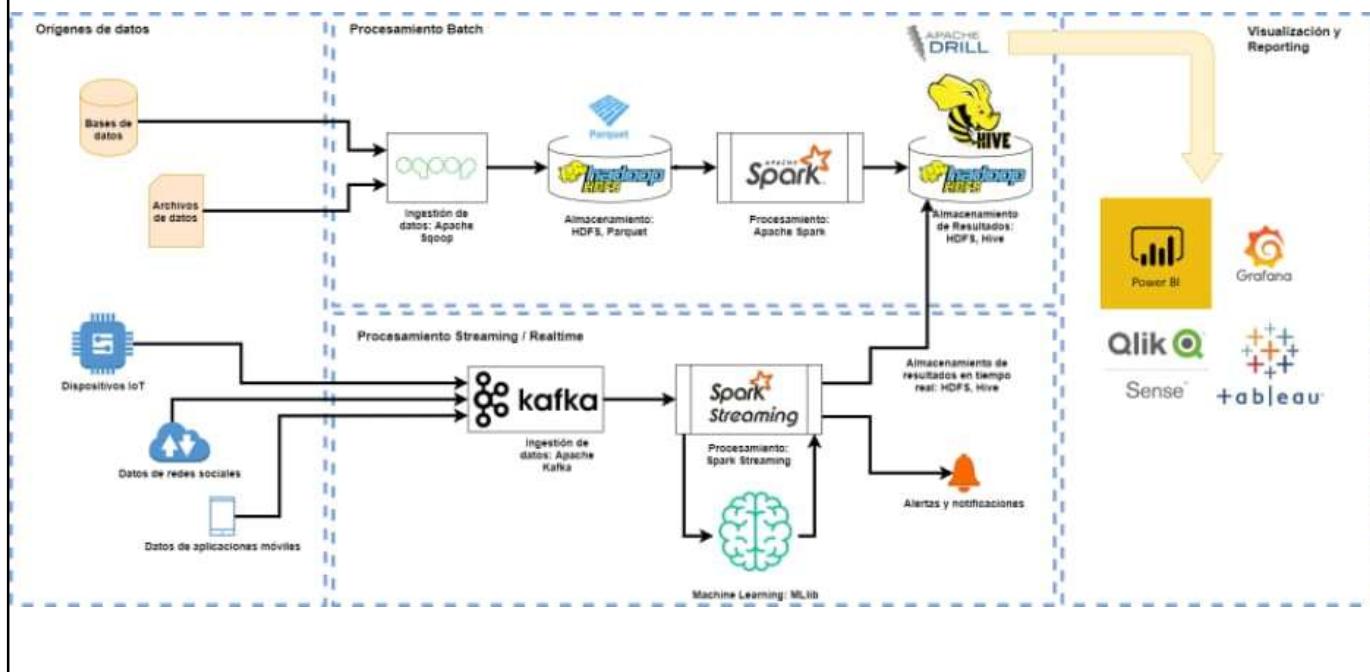
6

Diferentes puntos de vista para una misma cosa



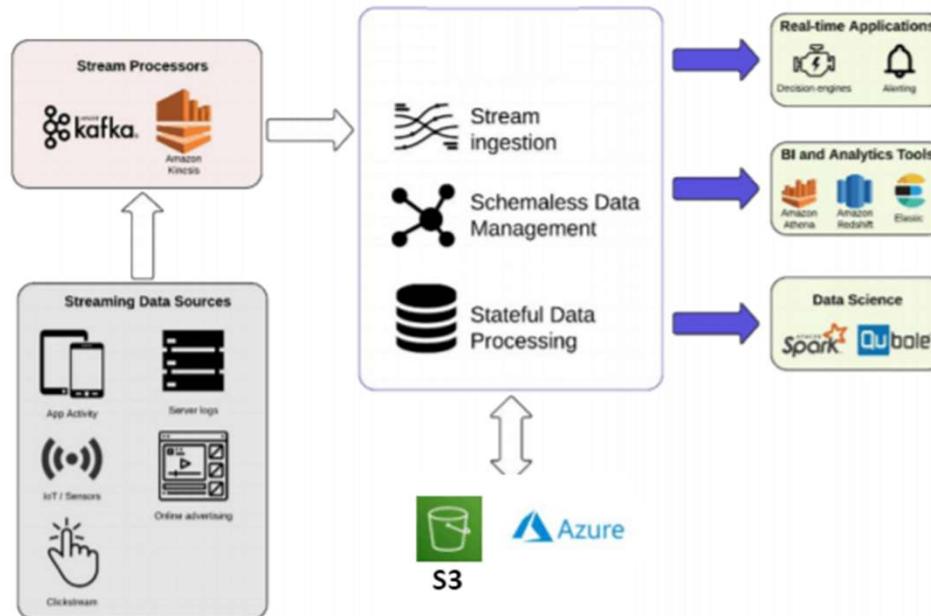
7

Ejemplo de arquitectura de datos de producción



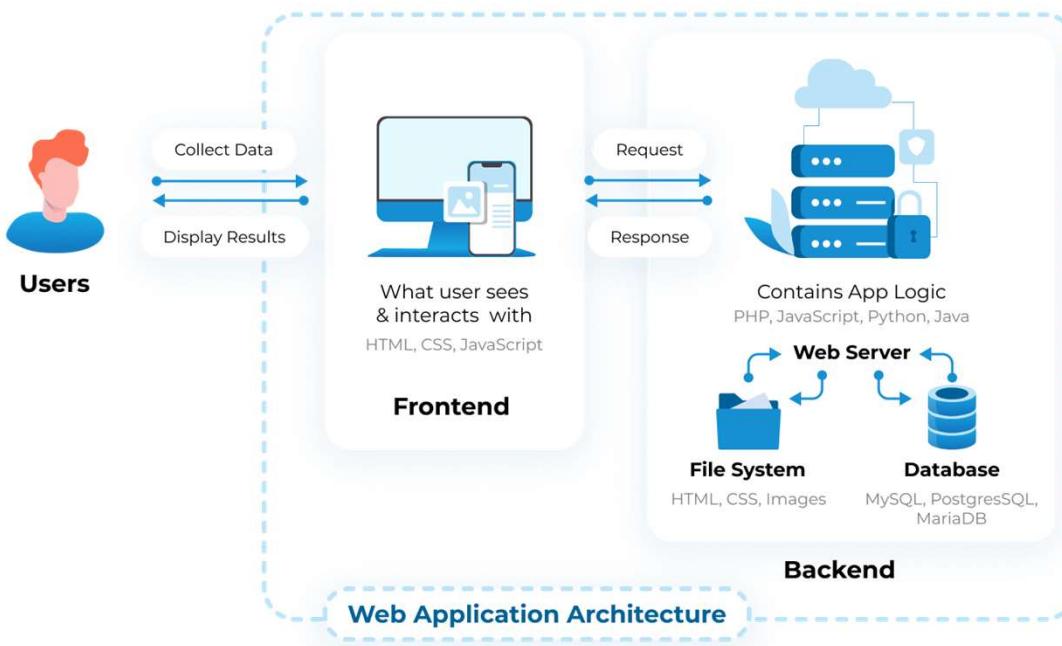
8

Ejemplo de arquitectura de tratamiento de datos en tiempo real

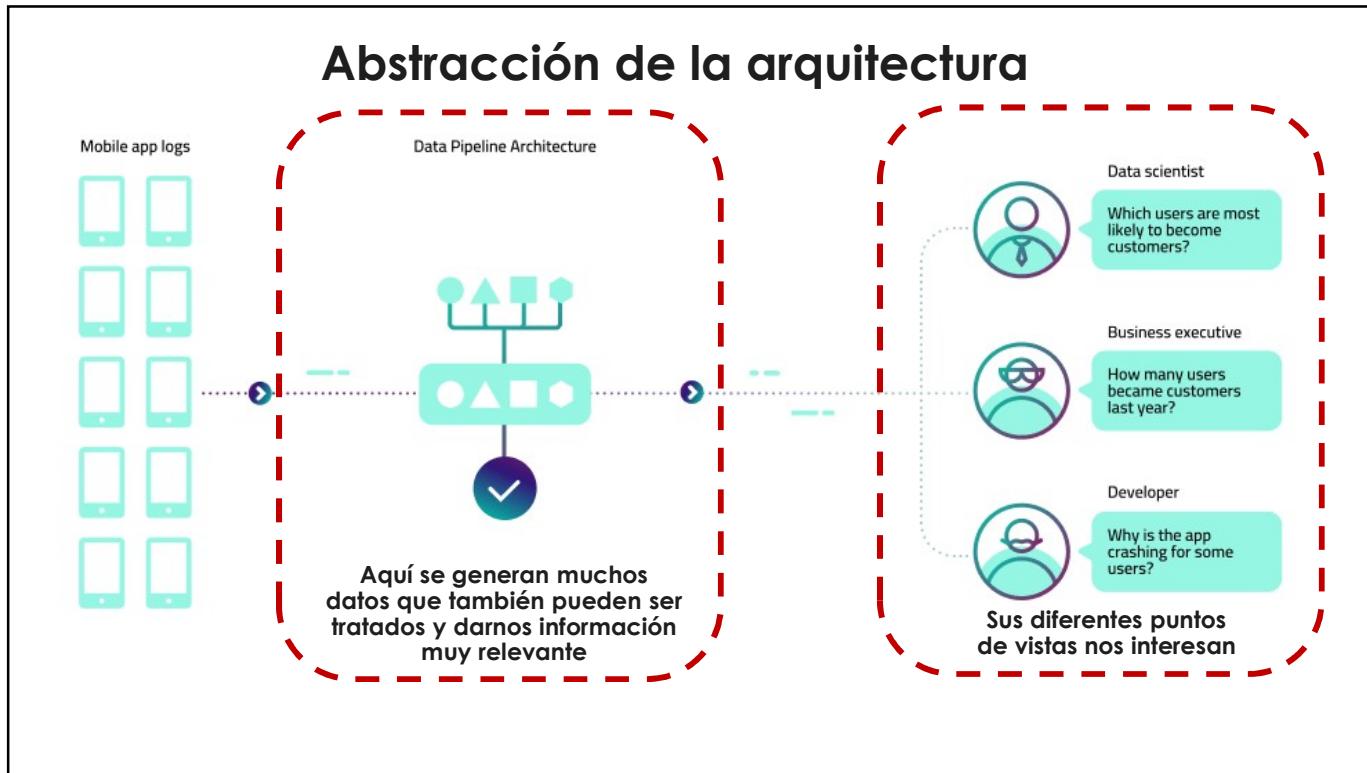


9

Ejemplo de arquitectura web/App



10

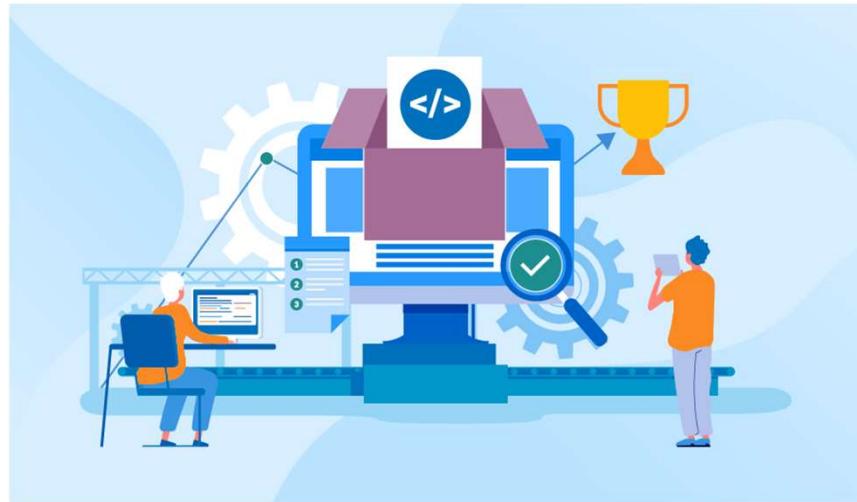


11



12

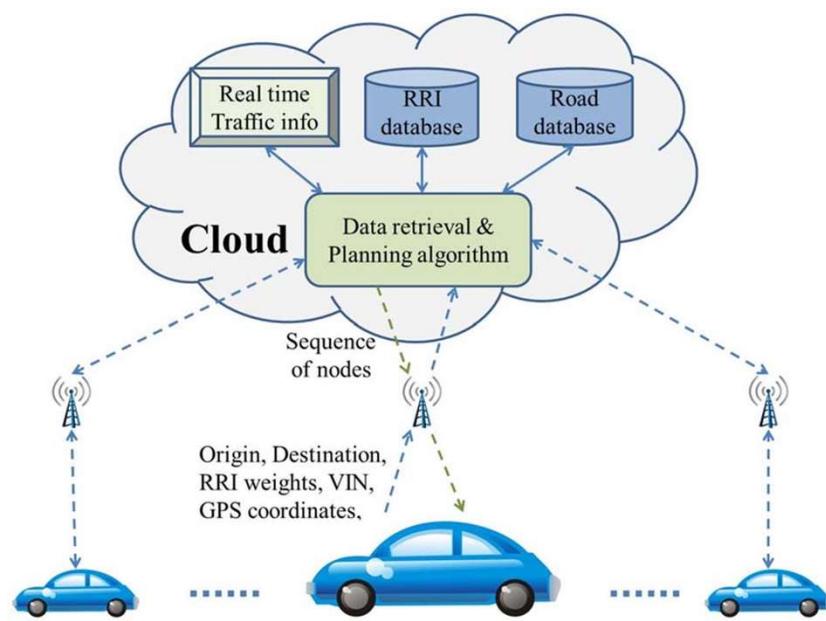
Mejorar la UX del Data Scientist



13

Optimizar el tráfico de datos a través de la arquitectura

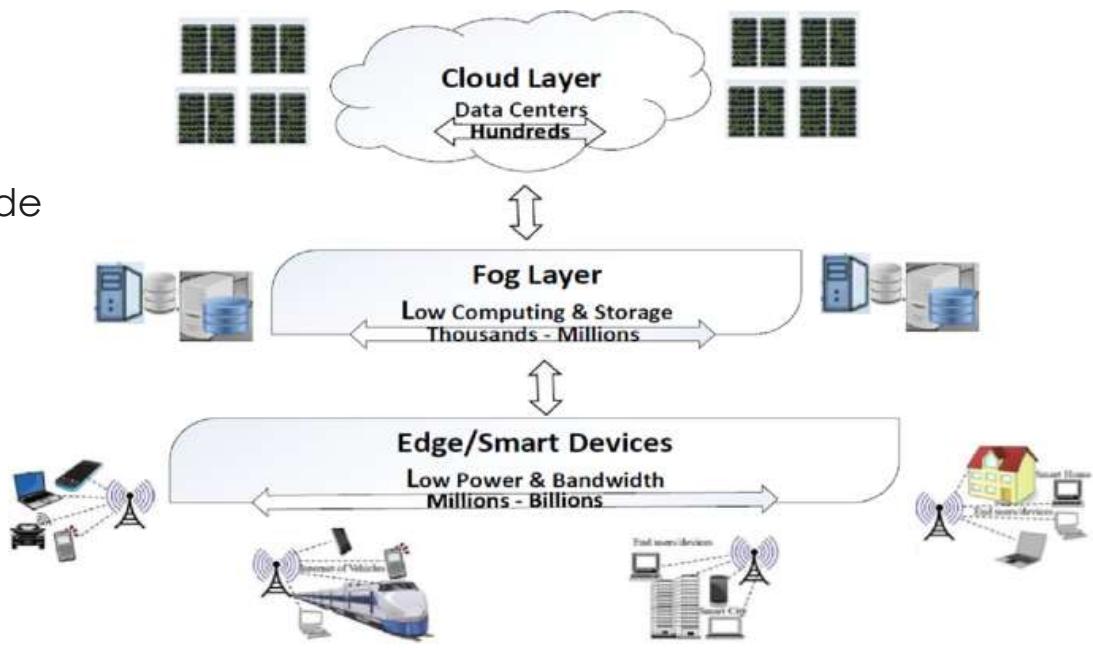
Detectar cuellos de botella



14

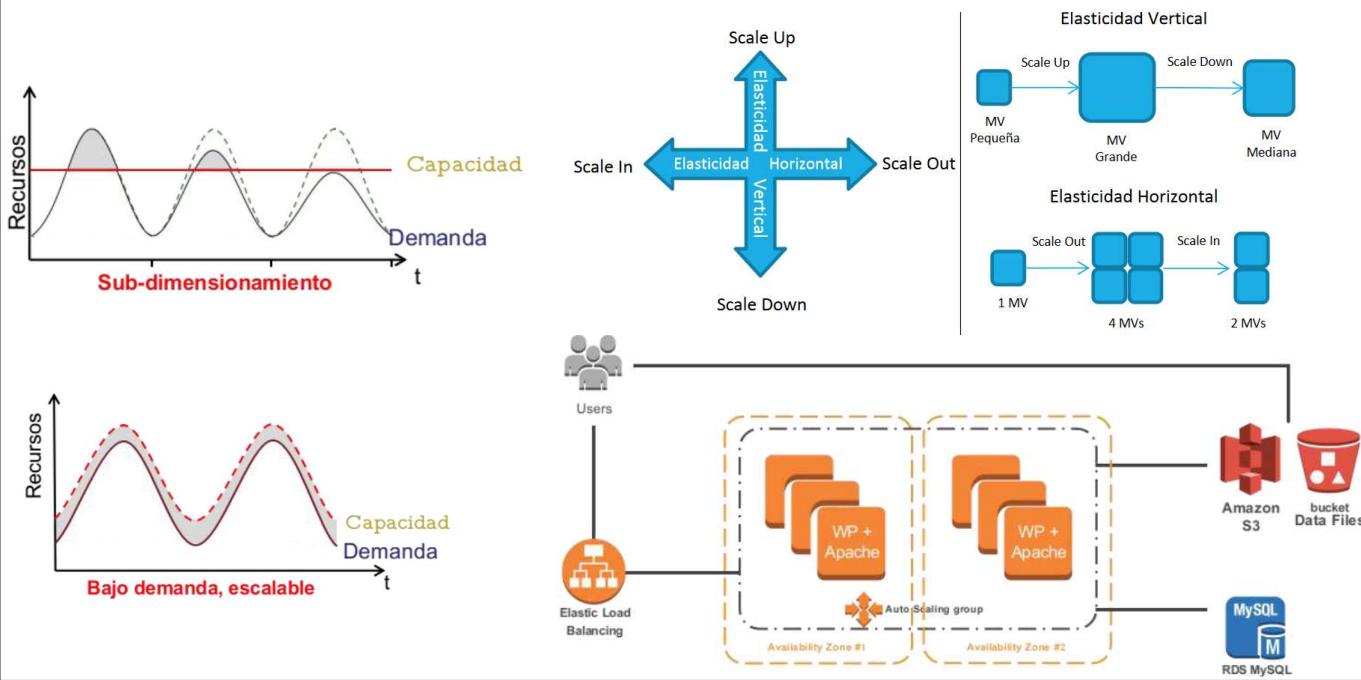
Optimizar la arquitectura

Evitar uso innecesario de recursos



15

Evitar caída de sistemas



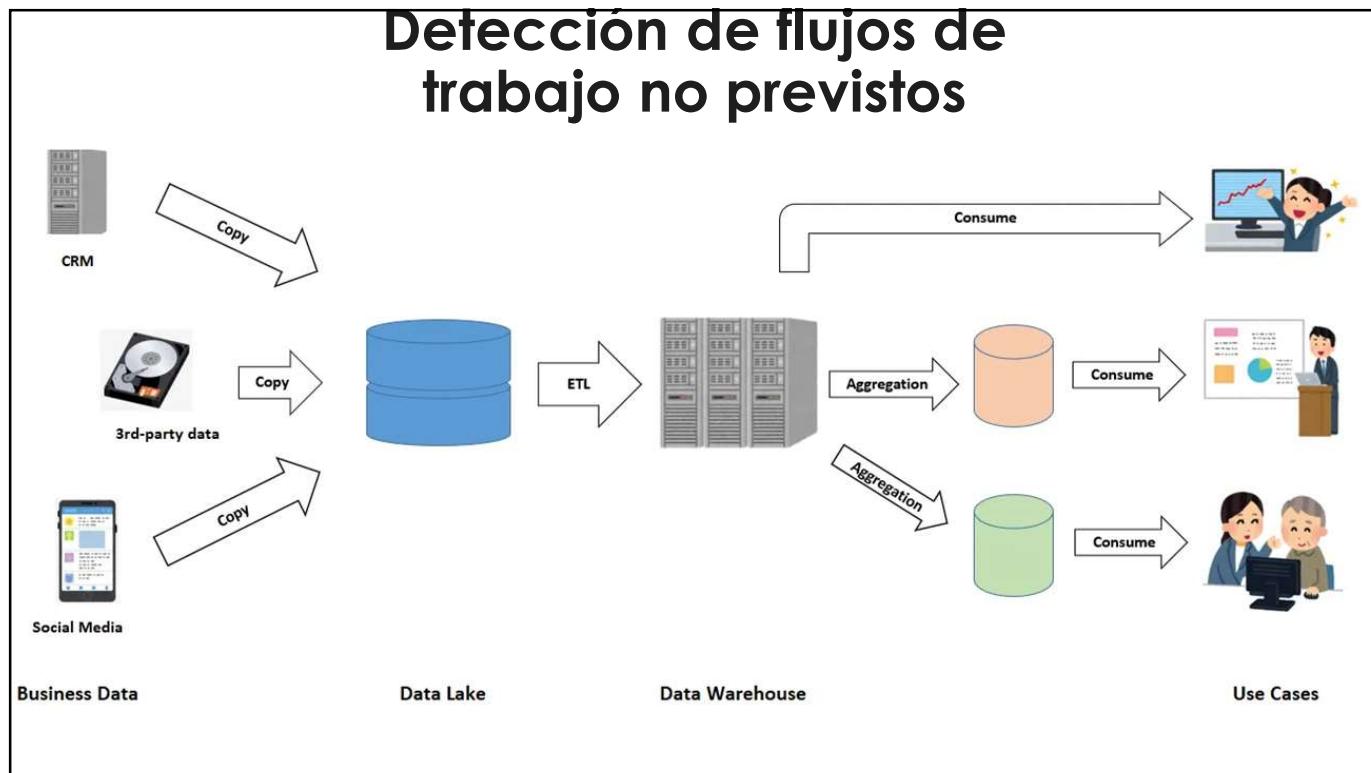
16

Detección de amenazas (ciberseguridad)



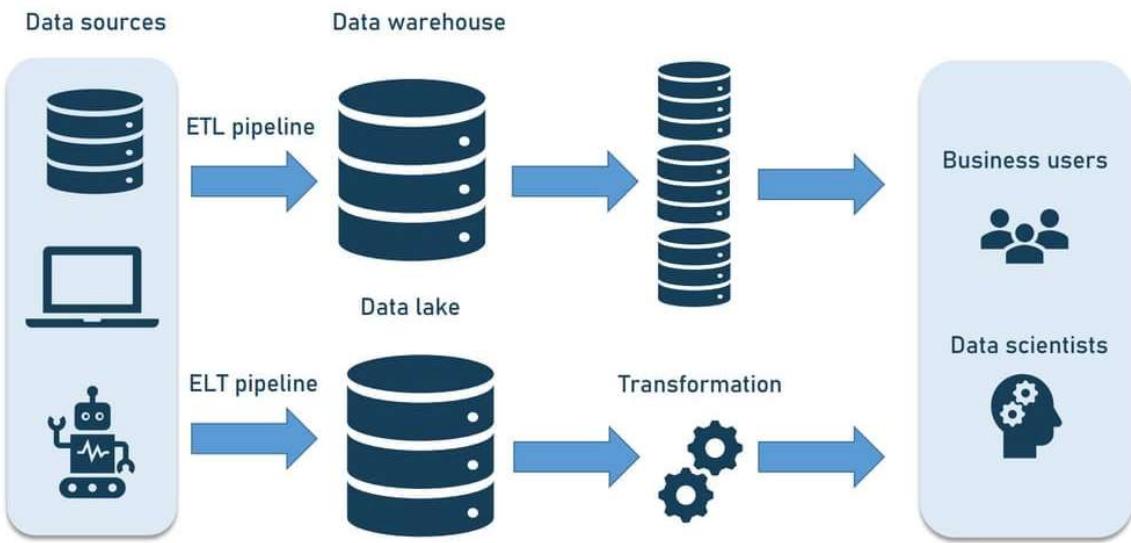
17

Detección de flujos de trabajo no previstos



18

Favorecer la accesibilidad



19

Flujo de trabajo

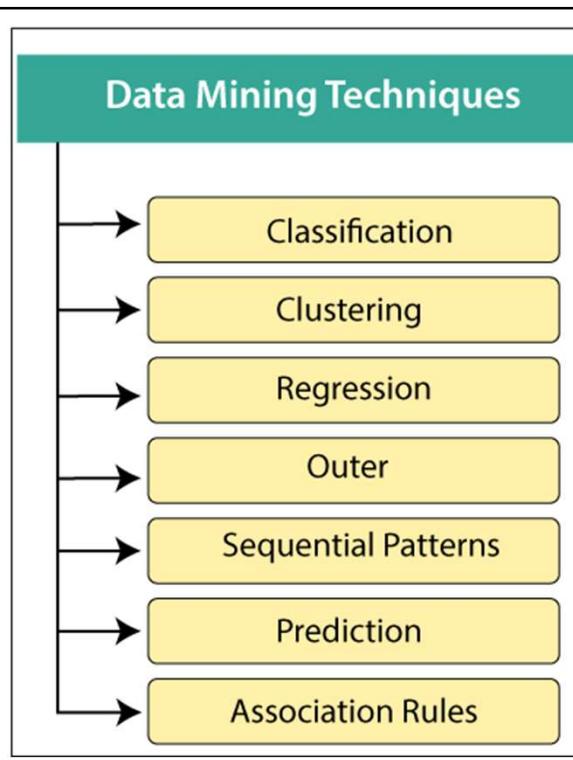
1. **KPI**
 2. **Ingesta**
 3. **Preparación**
 4. **Analisis**
 5. **Visualización**
 6. **Interpretacion**
1. **Establece los objetivos:** define qué quieres conseguir y establece unos KPI (Key Performance Indicator)
2. **Ingesta de datos:** identifica las fuentes de datos que están relacionadas con los objetivos marcados. Proceso ETL y ELT..
3. **Preparación de datos:** uso de técnica de limpieza y extracción de los KPI definidos.
4. **Análisis:** utilización de herramientas y modelos matemáticos para convertir datos en información (dashboards).
5. **Visualización de datos:** correcta representación de la información extraída de los datos (dashboards).
6. **Interpretación:** estudiar el análisis y representación realizados e interpretarlo.

20

Unidad 3 – Data Mining Techniques



1



2

CLASIFICACIÓN

Objetivo: establecer relaciones entre un conjunto de variables independientes con variables dependientes cualitativas.

- Obtiene información relevante clasificando los datos en diferentes categorías (clases).
- Utiliza métodos de aprendizaje supervisados.
- Dos tipos de modelos de clasificación: Generativos y Discriminativos.

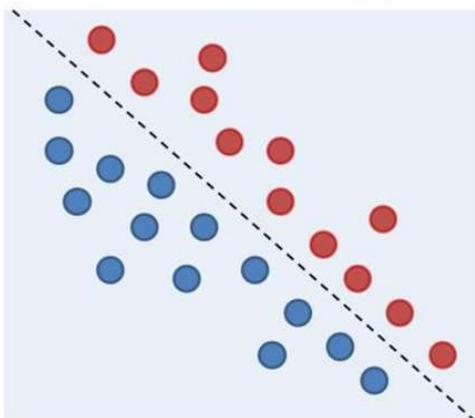
	Generativos	Discriminativos
¿Qué modela?	Cómo se generan (distribuyen) los datos	Cómo se separan los datos
Tipo de modelo	Solución única Función de distribución de probabilidad	Múltiples soluciones: Función discriminante
Clasificar una muestra nueva	Según probabilidad	Posicionamiento frente al discriminante
Principales desventajas	<ul style="list-style-type: none"> · Difícil saber la distribución real de las muestras · Afectado por outliers 	<ul style="list-style-type: none"> · Necesita histórico de datos relevante · Difícil elegir el mejor discriminante
Ejemplos	Clasificador de Bayes	Árbol clasificación, SVM...

3

CLASIFICACIÓN

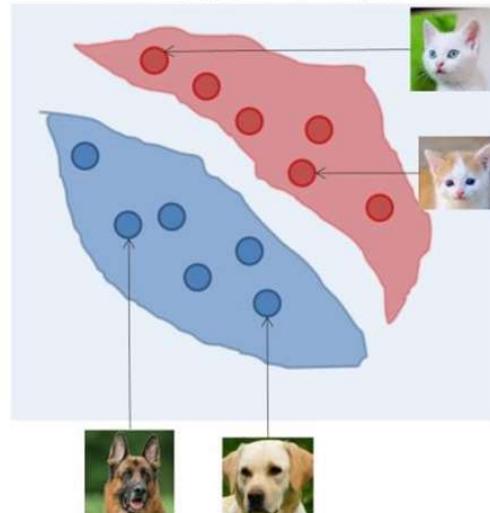
Discriminativo

Estimar directamente $P(y/x)$



Generativo

Estima $P(x/y)$ y deduce $P(y/x)$



X = características (variables del modelo)

Y = etiquetas (las diferentes clases)

4

¿Cómo podría usarse la técnica de clasificación en el Cloud Computing?

5

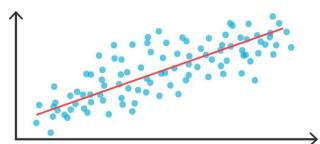
REGRESIÓN

Objetivo: Predecir el comportamiento de una variable dependiente, en función de un conjunto de variables independientes.

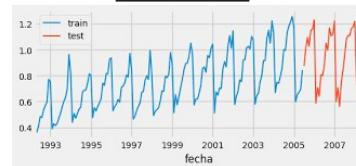
- Permite hacer predicciones del valor en función de unas variables conocidas.
- Utiliza métodos de aprendizaje supervisado, y puede ser lineal o no-lineal

Aplicaciones: principalmente para el modelado de datos y predicción.

Modelar

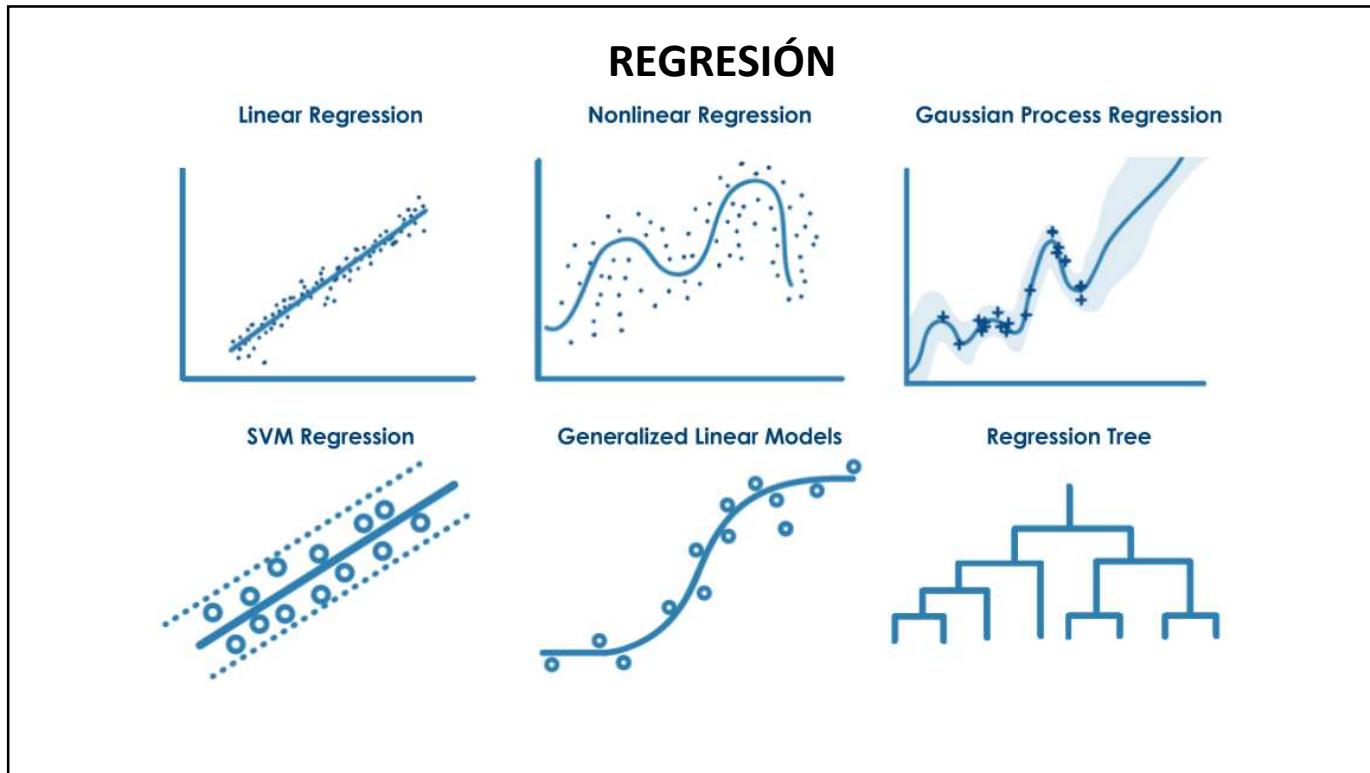


Predecir



Desventajas: Complejidad en su interpretación, así como en la vulnerabilidad frente a errores que puede afectar a su precisión.

6



7

**¿Cómo podría usarse la
técnica de regresión en
el Cloud Computing?**

8

PREDICCIÓN

Objetivo: descubrir patrones en datos históricos para proyectarlos a futuro.

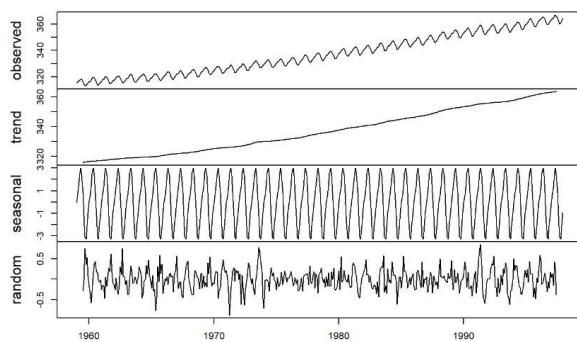
- Mezcla de Regresión + Análisis Estadístico.
- Aprendizaje supervisado + Análisis Estadístico.

Aplicaciones: principalmente para análisis de series temporales.

- Series temporales independientes o periódicas, de flujo (periodos largos) o de stock (intervalo concreto).
- Métodos cualitativos (no basados en históricos) y métodos cualitativos (se busca patrones en el pasado).
- Análisis de tendencias, Métodos de suavizado, ARIMA, Holt-Winters, aproximación por splines.
- Métodos de descomposición.

9

PREDICCIÓN



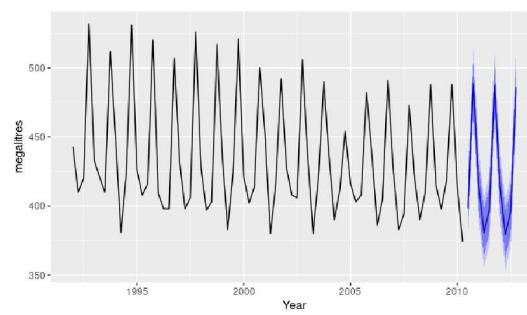
Componentes de Serie Temporal

Tendencia: evolución de la serie en el largo plazo

Fluctuación cíclica: variaciones periódica, no necesariamente regulares, a medio plazo en torno a la tendencia.

Variación Estacional: fluctuaciones regulares y repetitivas que a lo largo de un período de tiempo

Movimientos Irregulares: eventos aleatorios, identificables a posteriori



10

CLUSTERING

Objetivo: Conectar la información en grupos de elementos afines.

- Obtiene información a partir de las diferencias y similitudes entre datos → Se les asignan etiquetas.
- Utiliza métodos de aprendizaje no supervisado.

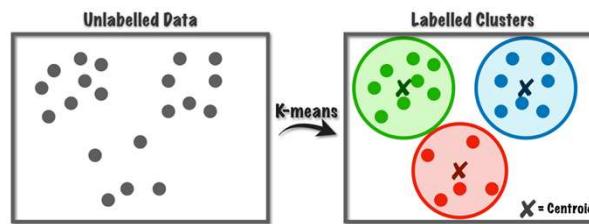
Aplicaciones: Análisis de redes, minado de textos, extracción de información relevante, análisis web, diagnosis médica, etc.

- Clustering clásico: semejanzas basadas en estadísticas.
- Clustering en ML: detección de patrones ocultos en las características de los datos.

11

CLUSTERING

	Clasificación	Clustering
Información a priori	Se conoce las clases de antemano	No se sabe cuántas clases (clusters) se obtendrán, ni qué características las definirán
Requiere etiquetas	Si, se necesitan grandes cantidades de muestras etiquetadas	No, se parte directamente de los datos existentes
Objetivo	Clasificar nuevas muestras en las clases conocidas a priori	Determinar la similitud entre datos
Ejemplos	Clasificador de Bayes, Regresión Logística, SVM...	K-means, mean-shift...



12

CLUSTERING

Técnicas que NO son clustering

Agrupar muestras por etiquetas predefinidas

Hacer divisiones o subgrupos de forma aleatoria

Agrupar muestras como resultado de queries a BBDD

13

OUTER

Objetivo: detectar muestras cuyas características no se corresponden con lo esperado (outlier analysis)

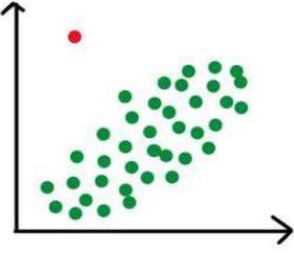
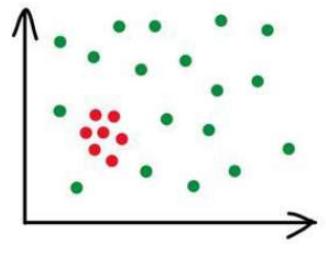
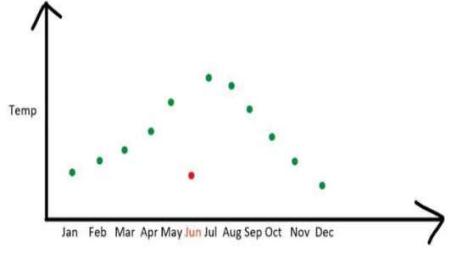
- También llamado outlier análisis
- Utiliza métodos de aprendizaje no supervisado, y puede ser lineal o no-lineal

Aplicaciones: limpieza de datos, detección de anomalías

- Detectar picos de venta fuera de lo común, detección de fraudes
- Comparativa: Tipos de outliers

14

OUTER

Outliers Globales (Puntuales)	Outliers Colectivos	Outliers Contextuales (Condicionales)
Muestras muy diferentes del resto · Detección de errores del sistema · Detección de Intrusos: detección puntual de un número elevado de paquetes enviados	Grupos de muestras colectivamente diferentes · Descubrimiento de comportamientos no deseados · Detección de Intrusos: detección de ataques DOS (Denial of Service)	Muestras normales en general, pero anómalas dentro de un contexto · Detección de operaciones sospechosas · Detección de intrusos: detección de envío de paquetes en horarios no previstos
		

15

REGLAS DE ASOCIACION

Objetivo: descubrir patrones de conexiones entre dos o más ítems o itemsets

- Mide la probabilidad de ciertas interacciones
- Tienen el formato de reglas tipo if-then: Si ocurre X, entonces ocurre Y → (X⇒Y)

Definiciones:

- Item: cada evento o elemento individual en una transacción
- Itemset: combinación específica de ítems
- Transacción: grupo de eventos asociados (cesta compra, historial de webs visitadas, etc).
 - Ejemplo: Items {A, B, C}, Transacción T = {A, B, C}, Itemsets posibles: {A,B,C}, {A,B}, {B,C}, {A,C}, {A}, {B} y {C}.

Aplicaciones: correlaciones de acciones, análisis de historiales médicos.

16

REGLAS DE ASOCIACION

Herramientas de medida de asociación:

	Descripción	Fórmula
SUPPORT (Soporte)	Frecuencia relativa: Frecuencia de un itemset {X,Y} en el total de transacciones	$S(\{X, Y\}) = \frac{\#Transacciones con \{X, Y\}}{\text{Total Transacciones}}$
CONFIDENCE (Confianza)	Probabilidad Condicionada: Cuando se produce el evento X, con qué frecuencia también se produce Y	$C(X \Rightarrow Y) = \frac{\#Transacciones con \{X, Y\}}{\#Transacciones con X}$
LIFT	Mejora de la Confianza: Mide si realmente hay una correlación entre los ítems de una confianza $C(X \Rightarrow Y)$. Cuando se da la asociación ($X \Rightarrow Y$), es porque X condiciona a que se de Y, o es simplemente que Y es frecuente de por sí.	$\text{Lift} = \frac{C(X \Rightarrow Y)}{S(\{Y\})}$

17

PATRONES DE SECUENCIAS

Objetivo: descubrir patrones secuenciales (subsecuencias de interés) dentro de datos secuenciales

- Subconjunto de las reglas de asociación incluyendo el elemento secuencial

Definiciones:

- Secuencia = conjunto de elementos ordenados. $S = \{e_1, e_2, \dots, e_N\}$, donde e_1 sucede antes de e_2 , etc.
- Elemento (transacción): muchas veces considerado como una sesión. Se compone de un conjunto de ítems
- Item (evento): componente básico de un elemento (el hecho a analizar)
- Secuencias de interés: según ocurrencia, frecuencia, longitud...

Aplicaciones: análisis de texto, análisis de la cesta de la compra (market basket analysis), navegación web

18

Unidad 4– Modelo de proceso orientado al análisis de datos



1

**¿Por qué este auge
de la ciencia de
datos?**

2



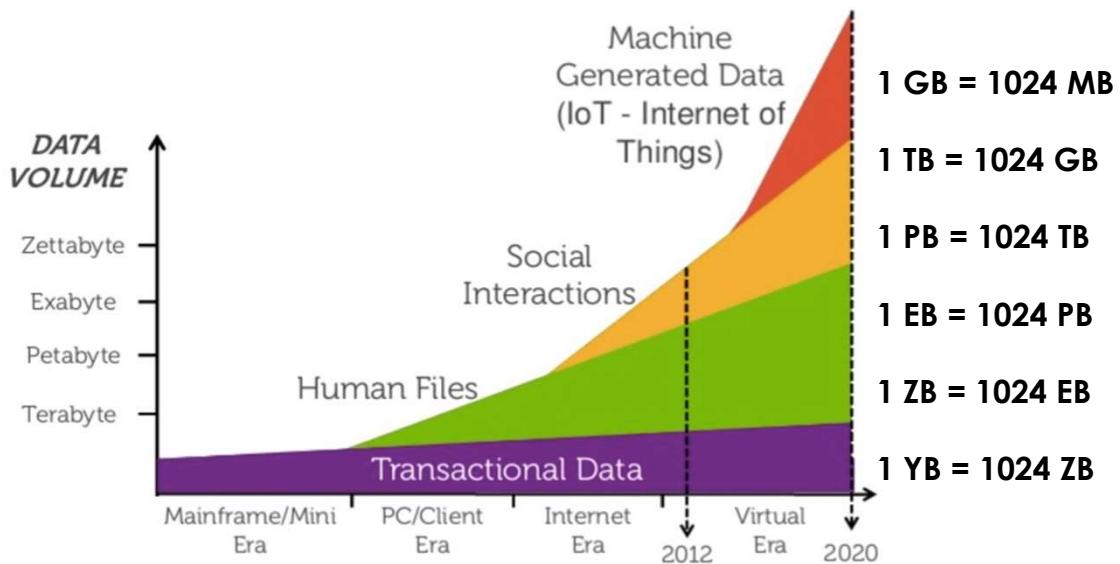
3

El cloud computing va de la mano de la ciencia de datos en su auge

4

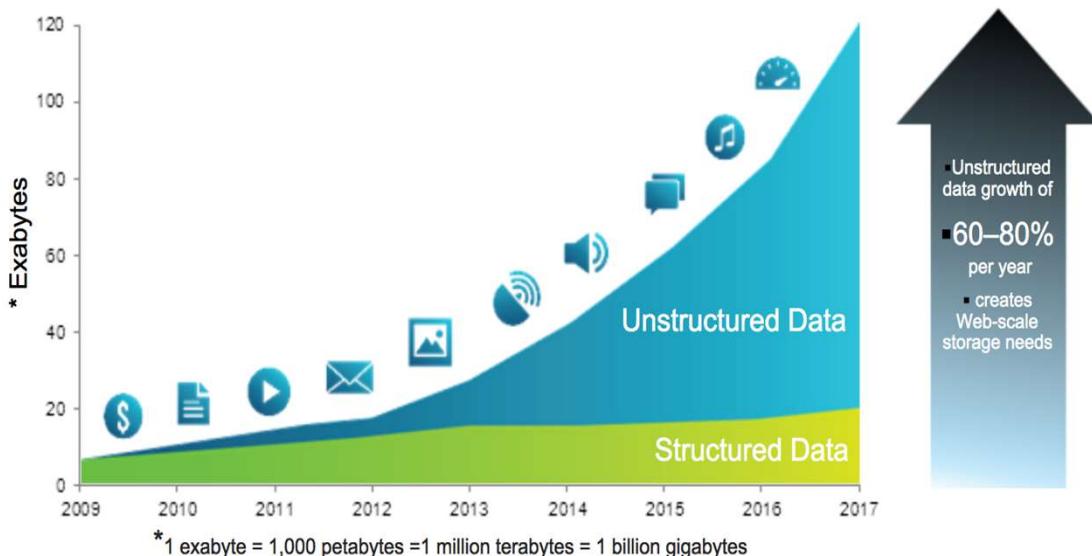
2

Algunos datos sobre datos



5

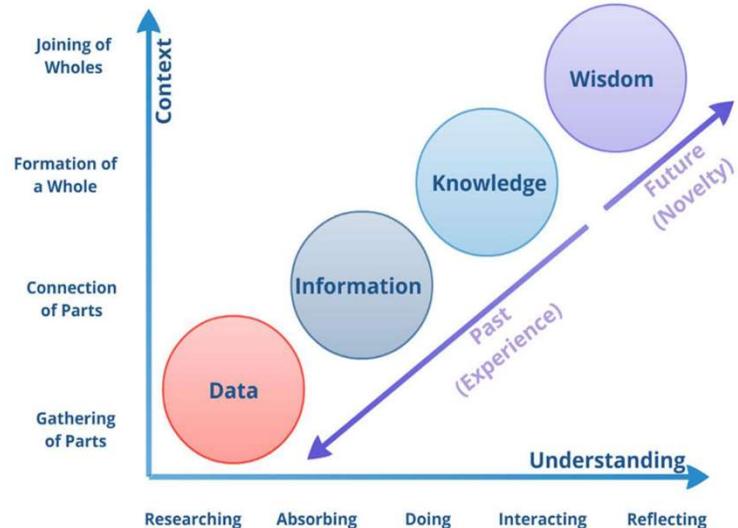
Algunos datos sobre datos



6

Beneficios del análisis de datos en el Cloud Computing

- Aprovechar información en toda su amplitud
- Proporciona respuestas
- Reducción de costes
- Eficiencia (mejor toma de decisiones)
- Rapidez en la ejecución
- Nuevos productos y oportunidades (relevancias ocultas)
- Reinención y creación de nuevos negocios
- Pro-actividad (Estrategias)



7

¿Qué significan las siglas KDD?

Knowledge Discovery in Databases

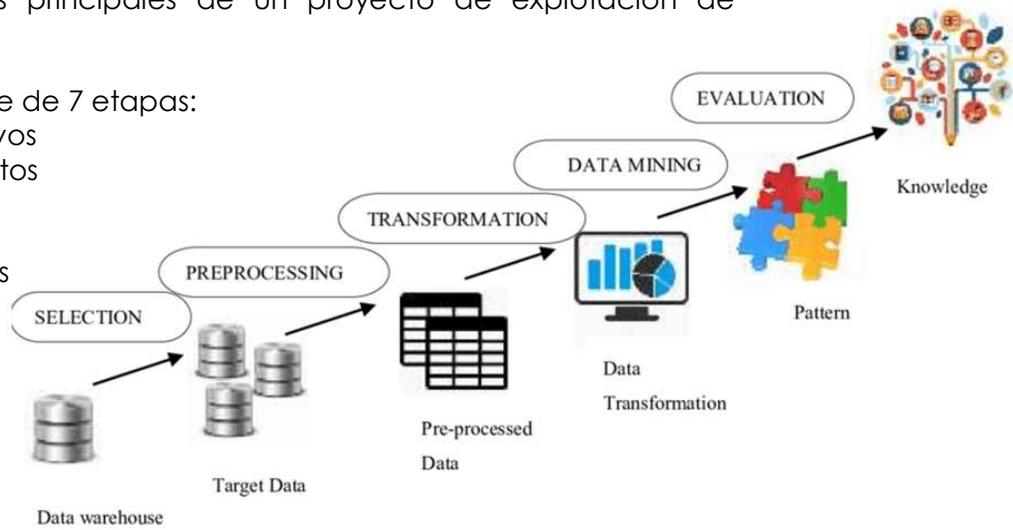
8

Knowledge Discovery in Databases

Primer modelo aceptado en la comunidad científica (1996) que establece las etapas principales de un proyecto de explotación de información.

Consta principalmente de 7 etapas:

- Selección objetivos
- Selección de datos
- Preprocesado
- Transformación
- Minado de datos
- Evaluación
- Conocimiento



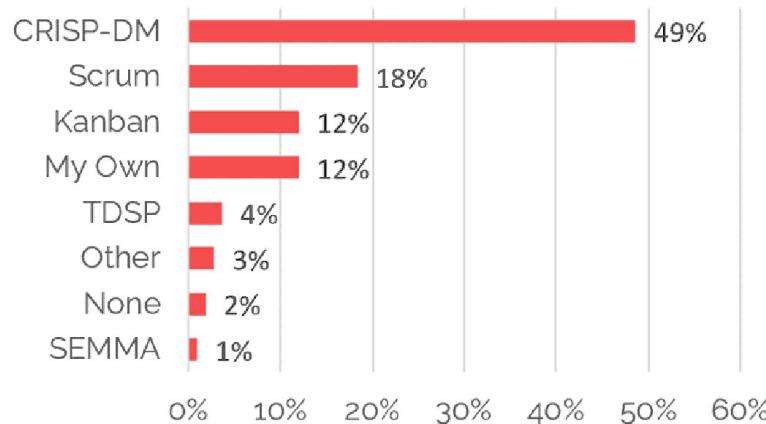
9

Knowledge Discovery in Databases

KDD es un **modelo iterativo e interactivo** para obtener conocimiento a partir de los datos

Es un **modelo genérico**, que requiere muchas **decisiones no automatizadas** y solo las **establece a grandes rasgos** (no profundiza en la descripción de tareas).

De este modelo surgen modelos más específicos.



10

¿Qué es el Data Mining?

11

Data Mining

Data Mining: Ciencia para explorar, extraer y explotar grandes conjuntos de datos (datasets).

- **Objetivo:** Extraer conocimiento inherente de los datasets, identificando patrones ocultos, generar predicciones, etc.
- **Dónde encontrar datos:** fuentes corporativas (bases de datos), sistemas IoT, web, social media, etc.
- Implica diferentes campos:
 - Gestión de BBDD y otras fuentes de datos
 - Estadística, matemática avanzada
 - Machine Learning
 - Deep Learning

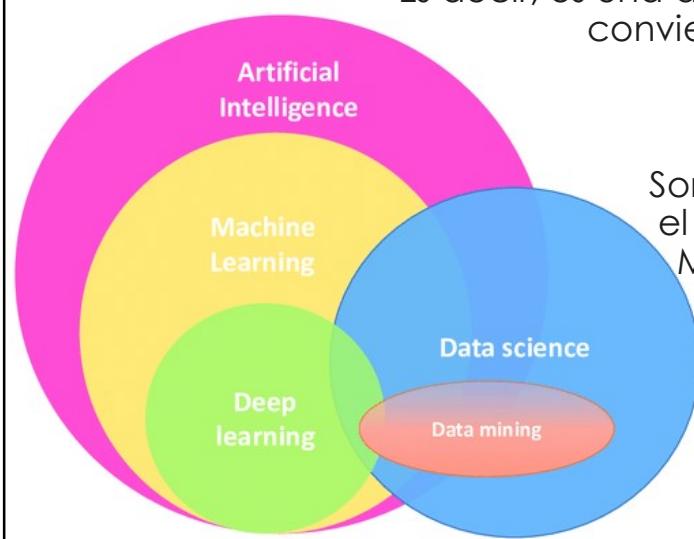
12

¿Data Mining es Data Science?

¿Data Mining es Machine Learning?

13

El Data Mining o Minería de Datos es una rama de la Data Science que nos permite derivar procesos automatizados o semiautomatizados para encontrar patrones o anomalías sobre Big Data, y explicar su comportamiento. Es decir, es una disciplina que hace que los datos se conviertan en información útil.



Son disciplinas interconectadas, donde el Data Mining se vale de procesos de Machine Learning para alcanzar sus objetivos.

14

CRISP-DM

Cross-Industry Standard Process for Data Mining



Fases para el Data Mining:

- **Entender el negocio:** Business Analyst. Objetivos de negocio a alcanzar (KPIs).
- **Entender los datos:** Data Analyst. Análisis de fuentes de datos, cantidad, calidad, propiedades, cómo nos ayuda a entender negocio.
- **Preparación de datos:** limpieza y almacenamiento.
- **Modelado:** Data Scientist. Detección de patrones, estadísticos, predicciones (machine learning).
- **Evaluación del modelo y resultados obtenidos.**
- **Visualización** (técnicas business intelligence).

15

Flujo de trabajo

1. KPI
 2. Ingesta
 3. Preparación
 4. Análisis
 5. Visualización
 6. Interpretación
1. **Establece los objetivos:** define qué quieres conseguir y establece unos KPI (Key Performance Indicator)
 2. **Ingesta de datos:** identifica las fuentes de datos que están relacionadas con los objetivos marcados. Proceso ETL y ELT..
 3. **Preparación de datos:** uso de técnica de limpieza y extracción de los KPI definidos.
 4. **Análisis:** utilización de herramientas y modelos matemáticos para convertir datos en información (dashboards).
 5. **Visualización de datos:** correcta representación de la información extraída de los datos (dashboards).
 6. **Interpretación:** estudiar el análisis y representación realizados e interpretarlo.

16

Gracias

- José María
Escalante
Fernández
- jm.escalante.fernandez@facultyue
.es

Unidad 5 – Analítica de LOGs



1

**¿Qué es un archivo
LOG?**

2

1

Archivos LOGs

Un archivo LOG es un evento que tuvo lugar en un momento determinado y puede tener metadatos que lo contextualicen.

Los archivos LOG son un registro histórico de todo lo que sucede dentro de un sistema, incluidos eventos como transacciones, errores e intrusiones.

Los datos en los archivos LOG pueden transmitirse de diferentes formas y pueden estar tanto en formato estructurado, semiestructurado como no estructurado.

```

216.239.46.60 - - [04/Jan/2003:14:56:50 +0200] "GET
/-lpis/curriculum/C+Unix/Ergastiria/Week-7/filetype.c.txt HTTP/1.0"
304 -
216.239.46.100 - - [04/Jan/2003:14:57:33 +0200] "GET
/~oswinds/top.html HTTP/1.0" 200 869
64.68.82.70 - - [04/Jan/2003:14:58:25 +0200] "GET /~lpis/systems/r-
device/r_device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/Jan/2003:14:58:27 +0200] "GET
/-lpis/publications/crc-chapter1.html HTTP/1.0" 304 -
209.237.238.161 - - [04/Jan/2003:14:59:11 +0200] "GET /robots.txt
HTTP/1.0" 404 276
209.237.238.161 - - [04/Jan/2003:14:59:12 +0200] "GET
/teachers/pitas1.html HTTP/1.0" 404 286
216.239.46.43 - - [04/Jan/2003:14:59:45 +0200] "GET
/~oswinds/publications.html HTTP/1.0" 200 48966

```

3

3

Anatomía del Archivo LOG

- **Marca de tiempo (time stamp):** hora a la que se produjo el evento registrado.
- **Información de usuario (user information):** información del usuario, software o máquina que generó ese archivo.
- **Información del evento (event information):** qué acción tuvo lugar.

4

4

Fuente de los Archivo LOG

- Aplicaciones
- Servidores
- Dispositivos IoT
- Aplicaciones
- Bases de datos
- Redes de comunicación
- ...



5

5

Tipología de los Archivo LOG

- **Event Log:** informacion sobre un evento determinado
- **Server Log:** informacion relacionada sobre la actividad de un servidor en un momento concreto
- **System Log (syslog):** registro temporal de los eventos de un sistema
- **Authorization Logs and Access Logs:** informacion relacionada con el acceso al sistema
- **Change Logs:** informacion cronologica de los cambios realizados en archivos o sistemas
- **Availability Logs:** seguimiento del rendimiento de sistemas
- **Resource Logs:** almacena informacion sobre la conectividad y capacidades del sistema
- **Threat Logs:** guarda informacion sobre actividades que el sistema considera poco usuales

6

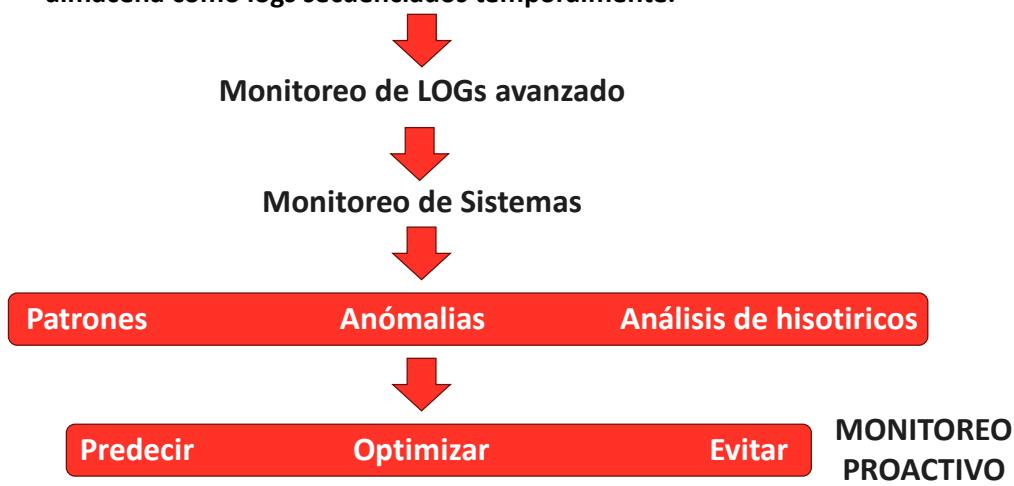
6

¿Qué es la Analítica de LOGs?

7

Analítica de LOGs

Las analíticas de logs son el proceso de búsqueda, investigación y visualización de datos generados por sistemas de TIC (Tecnologías de la Información y Comunicación), que se almacena como logs secuenciados temporalmente.



8

Como realizar la Analitica de LOGS

- Recopilar y centralizar datos
- Preparación y análisis de los datos
- Interpreta los datos
- Realizar cambios en consecuencia (optimización, monitoreo, alertas, ...)

9

Beneficios de la Analitica de LOGS

- Mejorar la experiencia de clientes
- Reducir el uso de recursos y la latencia del sistema
- Identificar comportamientos de usuarios
- Detectar actividades sospechosas
- Cumplimiento de normativas y estándares

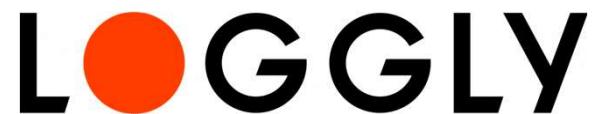
10

Desafios de la Analitica de LOGS

- Estandarización de LOGs
- Centralización de archivos
- Escalabilidad en la preparación y análisis de LOGs (crecimiento exponencial)

11

Herramientas para la Analitica de LOGS



12

Gracias

- José María
Escalante
Fernández
- jm.escalante.fernandez@facultyue
.es