

BIG DATA

- Material elaborado por el Profesor: Sergio Gevatschnaider



Photo by tian ci on Unsplash

INTRODUCCIÓN AL BIG DATA



Photo by Charlize on Unsplash

Introducción al Big Data

- **Definición de Big Data:** Conjunto de datos que por su volumen, velocidad y variedad requieren nuevas formas de procesamiento para mejorar la toma de decisiones.
- **Historia y evolución:** Desde los primeros sistemas de gestión de bases de datos hasta las modernas tecnologías de almacenamiento y procesamiento de grandes volúmenes de datos.
- **Importancia en el contexto actual:** El Big Data es crucial para la toma de decisiones en diversos sectores como salud, finanzas, marketing, etc.

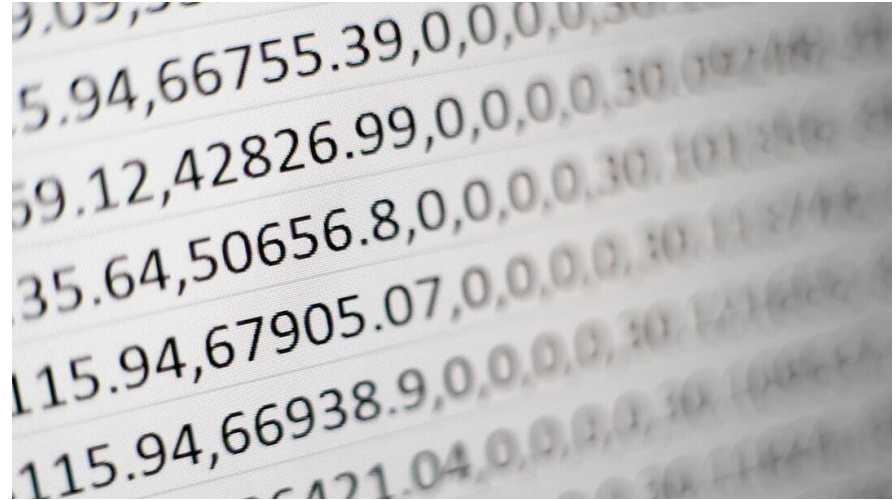


Photo by Mika Baumeister on Unsplash

Origen del Big Data

- **Antecedentes históricos:** El término 'Big Data' surge en los años 90, pero la gestión de grandes volúmenes de datos tiene sus raíces en los primeros sistemas de bases de datos.
- **Factores que impulsaron el crecimiento:** La digitalización masiva, el auge de Internet y las redes sociales, y los avances en tecnologías de almacenamiento y procesamiento.
- **Evolución de las tecnologías:** Desde los primeros data warehouses hasta las modernas plataformas de Big Data como Hadoop y Spark.



Photo by Museums Victoria on Unsplash

The Five V's of Big Data



Scale of Data

This refers to the sheer volume of data being generated every second.

6 Billion People have cell phones

40 Zettabytes of data will be created by 2020 and increase of 300 times from 2005

Most companies in the U.S. have at least **100 Terabytes** of data stored.



Uncertainty Of Data

1 in 3 Business leaders don't trust the information they use to make decisions

This refers to the discrepancies found in the data.

Poor data quality costs the US economy around **\$ 3.1 Trillion a year**



Analysis of Streaming Data

The New York Stock Exchange capture **1 TB of Trade Information**

Denotes the speed at which data is emanating and changes are occurring between the diverse data sets.



By 2016 it is projected there will be **18.9 Billion** network connections

Modern cars have close to **100 Sensors**



4 Billion+ hours of video are watched on YouTube each month

30 Billion pieces of content are shared on Facebook every month

400 Million tweets are sent per day by about 200 million monthly active users

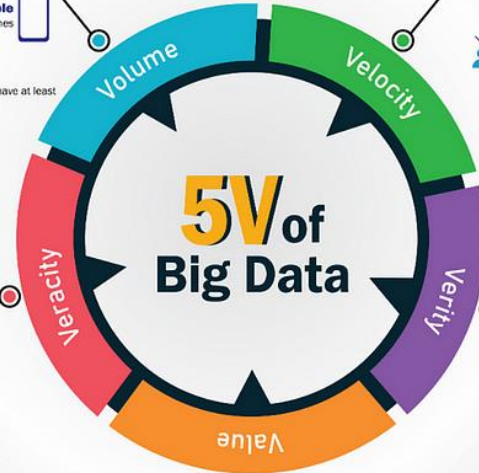
Different forms of data

As more and more data is being digitized.



Value Of Data

Having access to big data is all well and good but that's only useful if we can turn it into a value.



POPULAR BIG DATA USE CASES BY INDUSTRY

FINANCE

Credit Scoring
Fraud Prevention

HEALTHCARE

Predictive Analytics
Prescriptive Analytics

TRANSPORTATION

Autonomous Vehicles

RETAIL

Customer Service
Inventory Positioning
Cart Size Optimization

MEDIA & ENTERTAINMENT

Recommender Systems

FMCG

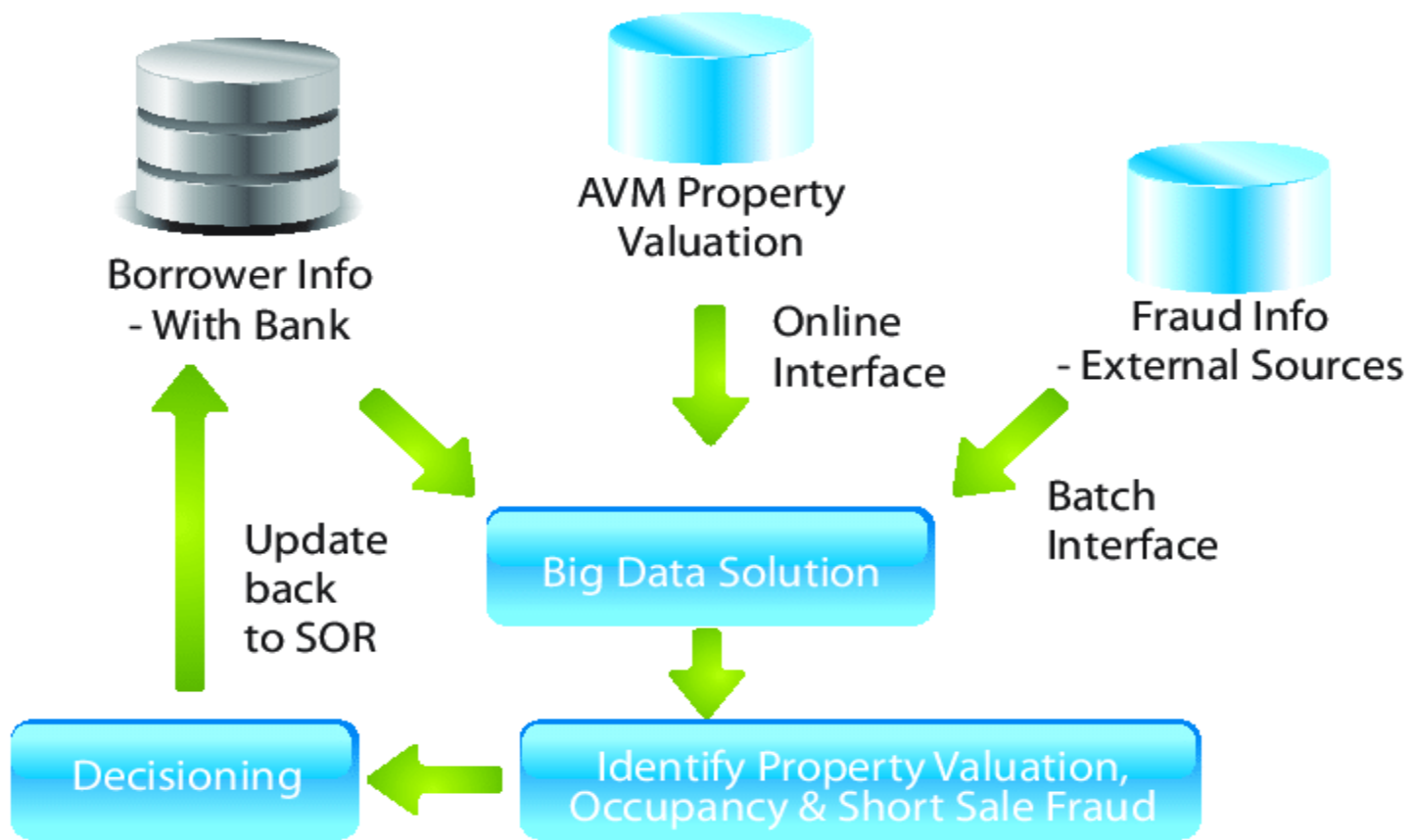
Lean Supply Chain Optimization
Waste Minimization

MANUFACTURING

Predictive Maintenance

WAREHOUSING

Inventory Positioning



BIG DATA USE CASES IN THE HEALTHCARE INDUSTRY



Here are five use cases of big data when implemented well in the healthcare industry:



**Predictive
Analytics for
Patient Outcomes**



**Facilitating
medical
research**



**Enabling
real-time
alerts**

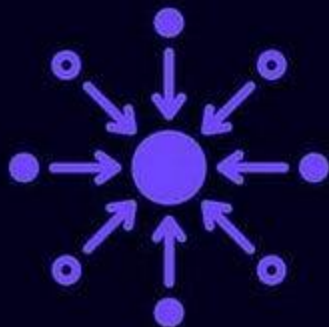


**Preventing
cyberattacks
and fraud**



**Telemedicine
and virtual
consultations**

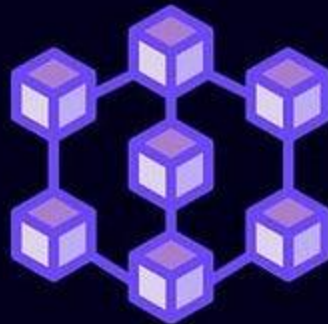
Web 1.0



Web 2.0



Web 3.0

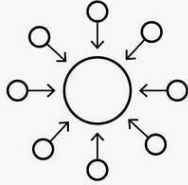


Evolution of the web from 1.0 to 3.0



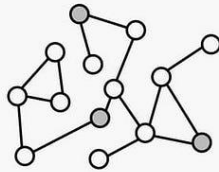
1900s–2000

Static read-only
web pages



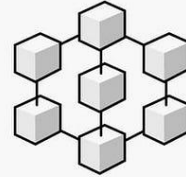
2000s–2020s

Information-centric
and interactive



2020s–?

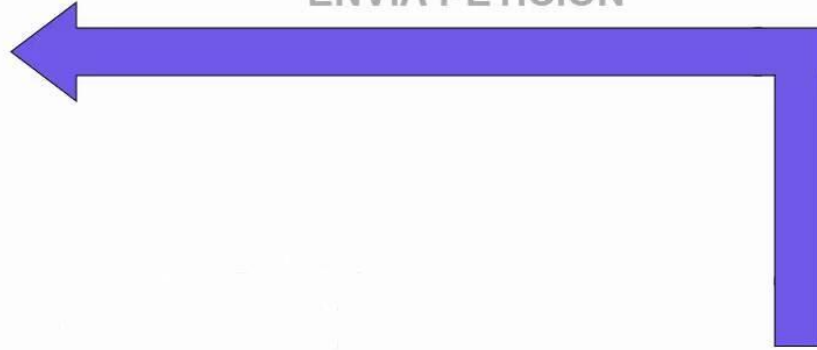
User-centric,
decentralized, private,
and secure





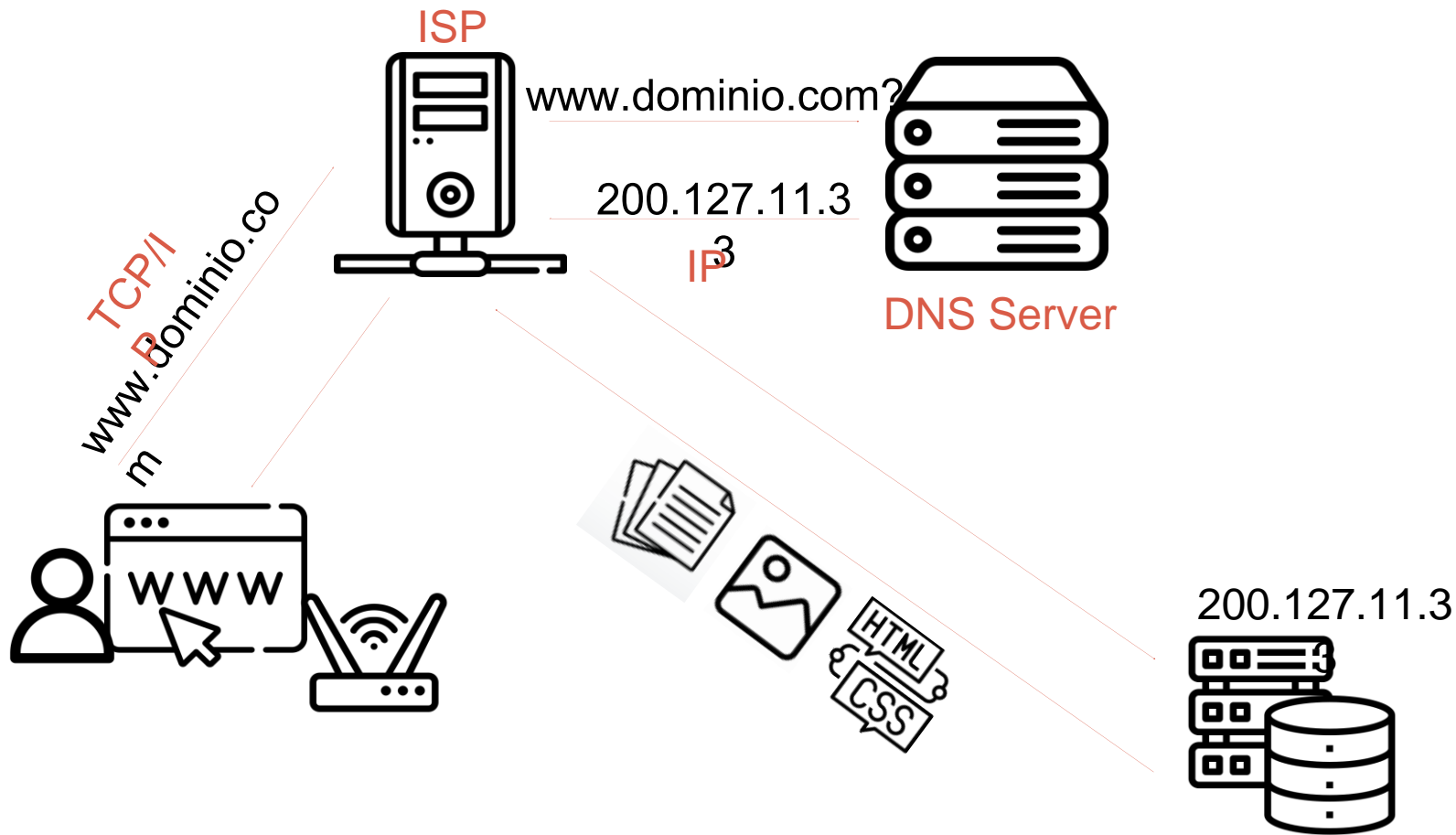
SERVIDOR

ENVIA PETICION



CLIENTE



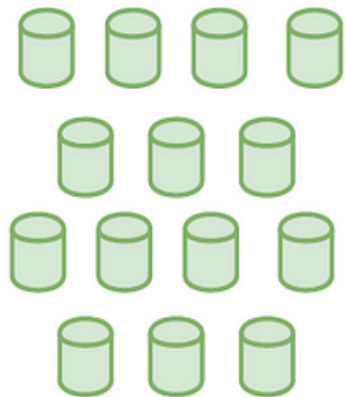


192.168.1.1

2001:0db8:85a3:0000:0000:8a2e:0370:7334

2001:db8:85a3::8a2e:370:7334

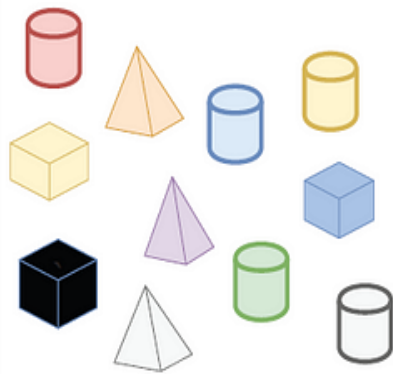
Data at rest



Terabytes to zettabytes
of data to process

Volume

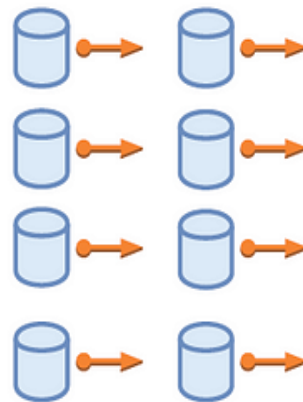
Data in many forms



Structured,
unstructured, and semi-
structured

Variety

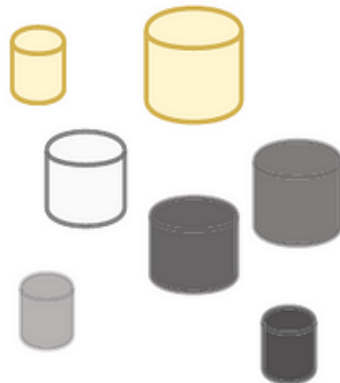
Data in motion



Streaming data,
microseconds to seconds
to respond

Velocity

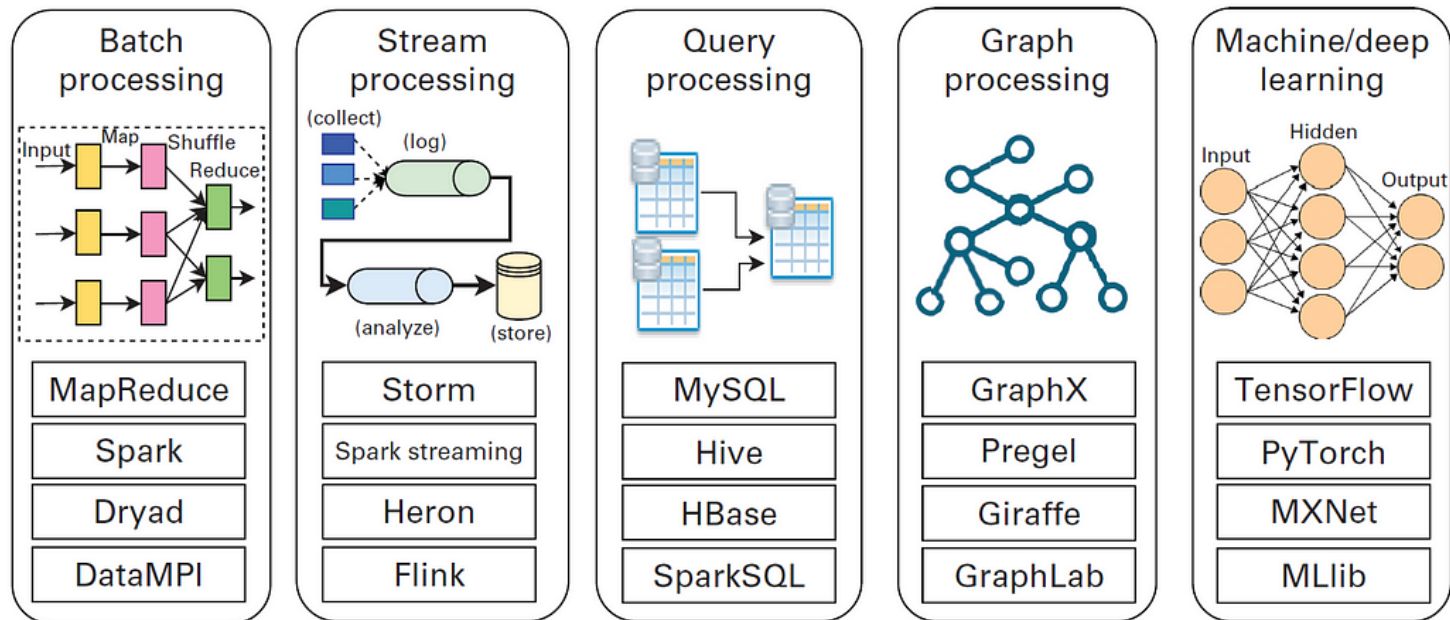
Data in doubt



Uncertainty due to data
inconsistency, ambiguities,
deception, and model
approximations

Veracity

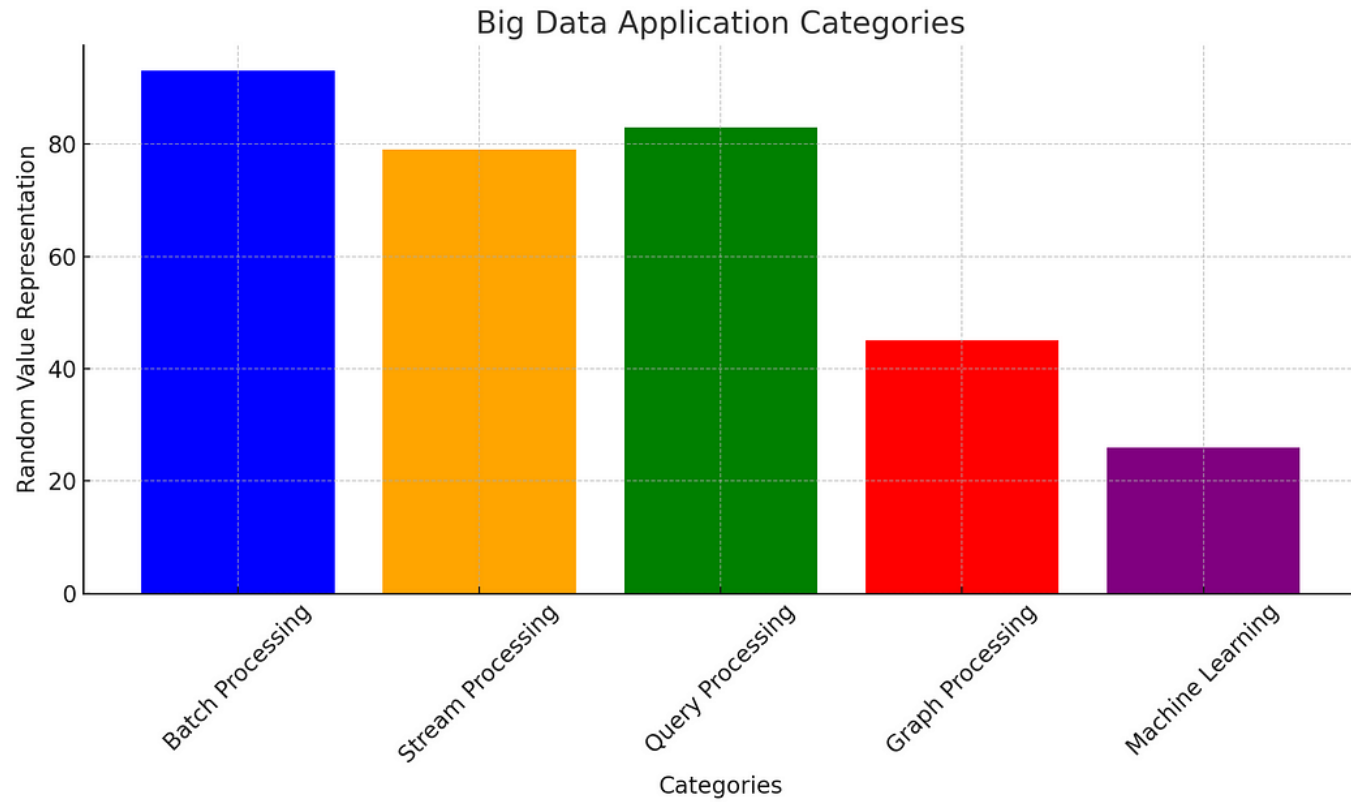
Big data applications



Storage middleware and API interface

Escala Completa de Unidades de Almacenamiento de Datos:

- Kilobyte (KB): 10^3 bytes (1,024 bytes en binario).
- Megabyte (MB): 10^6 bytes (1,048,576 bytes en binario).
- Gigabyte (GB): 10^9 bytes (1,073,741,824 bytes en binario).
- Terabyte (TB): 10^{12} bytes (1,099,511,627,776 bytes en binario).
- Petabyte (PB): 10^{15} bytes (1,125,899,906,842,624 bytes en binario).
- Exabyte (EB): 10^{18} bytes (1,152,921,504,606,846,976 bytes en binario).
- Zettabyte (ZB): 10^{21} bytes (1,180,591,620,717,411,303,424 bytes en binario).
- Yottabyte (YB): 10^{24} bytes (1,208,925,819,614,629,174,706,176 bytes en binario).



Structured Data



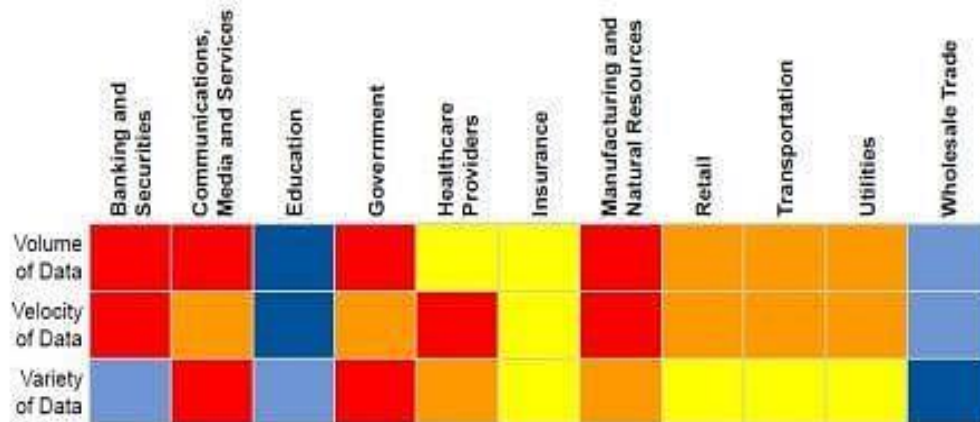
Unstructured Data



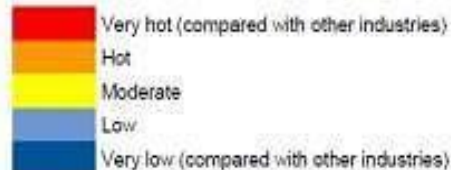
Semi-structured Data



Comparison of Data Characteristics by Industry



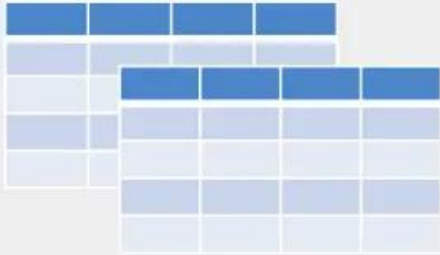
Potential big data opportunity on each dimension is:



BBDD Relacionales y no relacionales

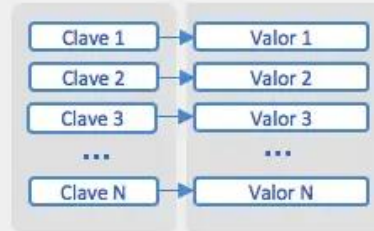
Base de datos SQL

Relacional



Base de datos NoSQL

Clave-Valor



Documental



Grafos

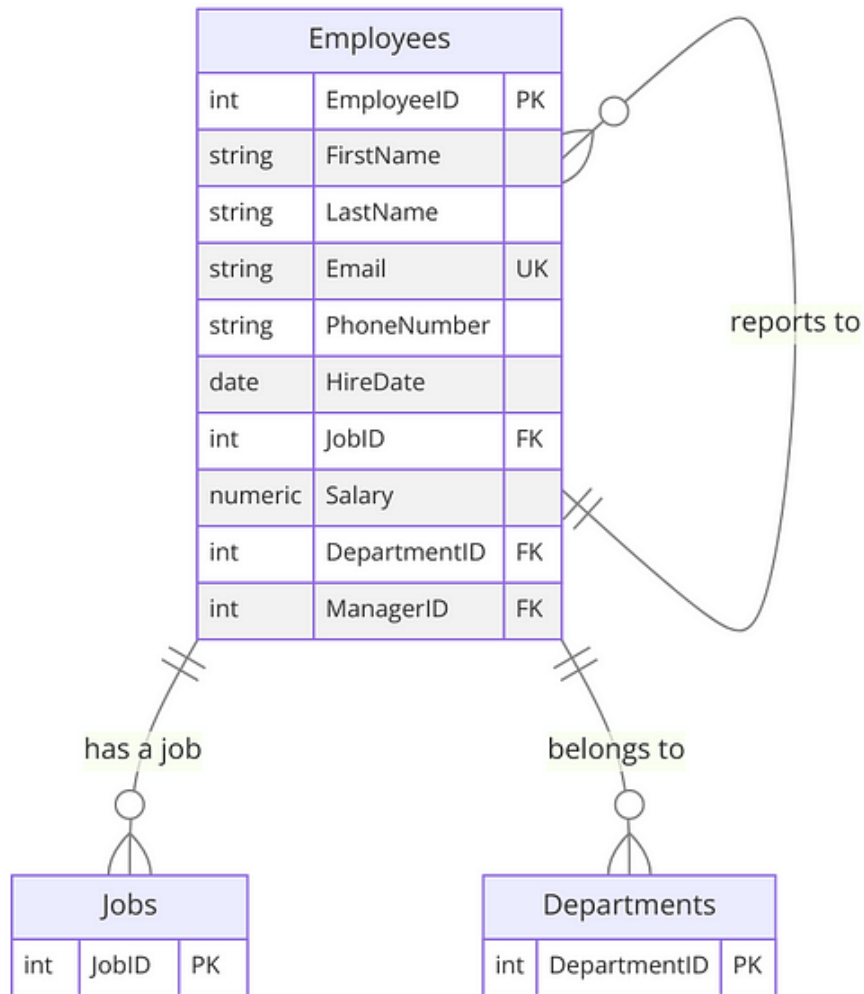


BBDD Relacionales y no relacionales

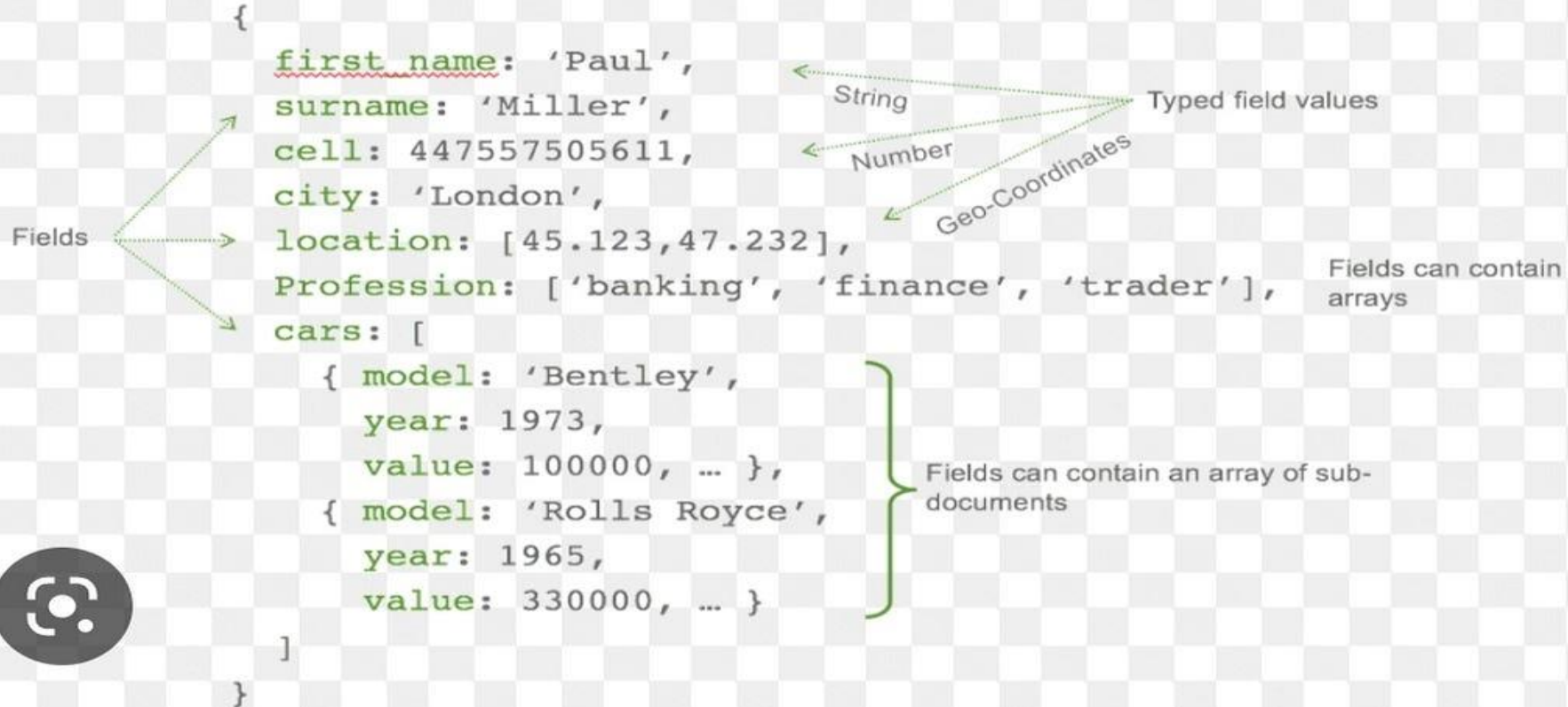
Relacionales	No relacionales
Tienen un identificador que establecen una relación entre dos grupos de datos	No tienen un identificador que sirva de relación entre un conjunto de datos y otros
La información se organiza en tablas	La información se organiza en documentos
	Se utiliza cuando no tenemos un esquema exacto de lo que se va a almacenar

BBDD Relacionales (SQL server, Oracle, etc)

```
CREATE TABLE Employees (  
  EmployeeID SERIAL PRIMARY KEY, -- Auto-incrementing ID for each  
  employee, serving as the primary key  
  FirstName VARCHAR(50), -- Employee's first name  
  LastName VARCHAR(50), -- Employee's last name  
  Email VARCHAR(100) UNIQUE, -- Employee's email address, must be  
  unique  
  PhoneNumber VARCHAR(15), -- Employee's phone number  
  HireDate DATE, -- Date the employee was hired  
  JobID INTEGER REFERENCES Jobs(JobID), -- Foreign key referencing the  
  Jobs table  
  Salary NUMERIC(10, 2), -- Employee's salary stored as a numeric type for  
  precise decimal calculations  
  DepartmentID INTEGER REFERENCES Departments(DepartmentID), --  
  Foreign key to the Departments table  
  ManagerID INTEGER REFERENCES Employees(EmployeeID) -- Foreign key  
  pointing to another employee who is the manager  
);
```



MONGO DB



BBDD Relacionales y no relacionales

```
SELECT E.FirstName, E.LastName, E.Email, J.JobTitle,  
D.DepartmentName  
FROM Employees E  
JOIN Jobs J ON E.JobID = J.JobID  
JOIN Departments D ON E.DepartmentID = D.DepartmentID  
WHERE E.Salary > 50000;
```

NoSQL Databases (e.g., MongoDB)

```
{  
  "employee_id": "12345",  
  "first_name": "Pedro",  
  "last_name": "Perez",  
  "date_of_birth": "1988-01-18",  
  "address": {  
    "city": "City",  
    "state": "BA",  
    "zip_code": "12"  
  },  
  "hire_date": "2012-06-01",  
  "manager_id": "1000"  
}
```


employee_id	12345
first_name	Pedro
last_name	Perez
date_of_birth	1988-01-18
address	●
hire_date	2012-06-01
manager_id	1000

city	City
state	BA
zip_code	12

db.employees.find({ "employee_id": "12345" })

RDBMS vs MongoDB

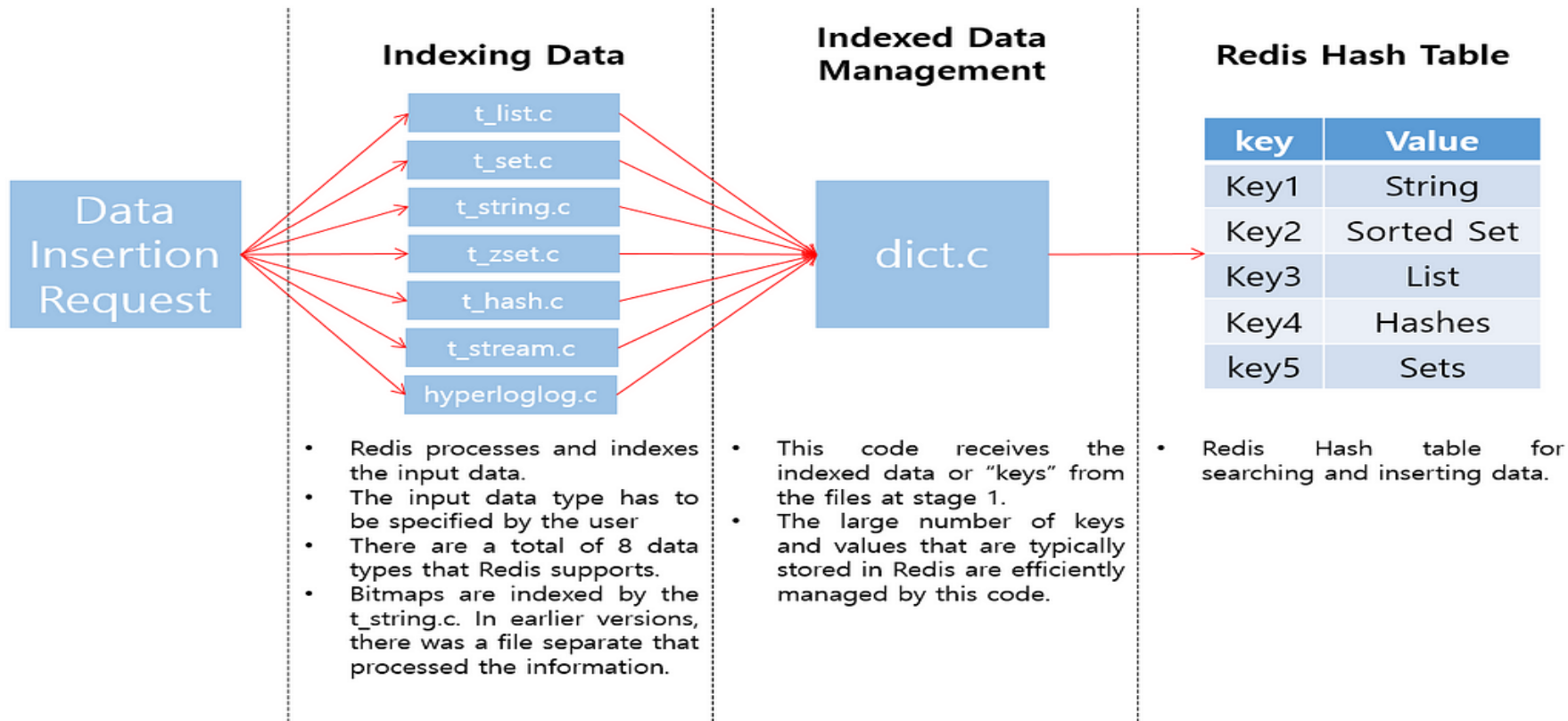
Comparison Chart

RDBMS	MongoDB
RDBMS is a relational database model in which data is stored in multiple tables.	MongoDB is an open-source, document-oriented database that has no concept of tables, schemas, rows, or SQL.
Records are stored as rows in tables, wherein table are organized into columns with each column attributed to one data type.	MongoDB uses different formats to store data such as document stores, graph databases, key-value stores, and more.
It follows a typical schema design comprises of several tables and relationships between them.	It is based on a schema-less data representation with no regards to the concept of relationship.
RDBMS databases are vertically scalable meaning when database loads increase, you scale database by increasing the capacity of existing hardware.	MongoDB is a one-size-fits-all database and is considered to be more scalable than the traditional RDBMS database models.
	

Key-Value Stores (e.g., Redis)

```
SET employee:12345 '{"employee_id": "12345",  
  "first_name": "Pedro", "last_name": "Perez",  
  "date_of_birth": "1988-01-18", "address": {"city": "City",  
  "state": "BA", "zip_code": "12"}, "hire_date": "2012-06-  
01", "manager_id": "1000"}'
```


Key-Value Stores (e.g., Redis)



BBDD no relacionales

orientadas a columnas

Base de datos relacional

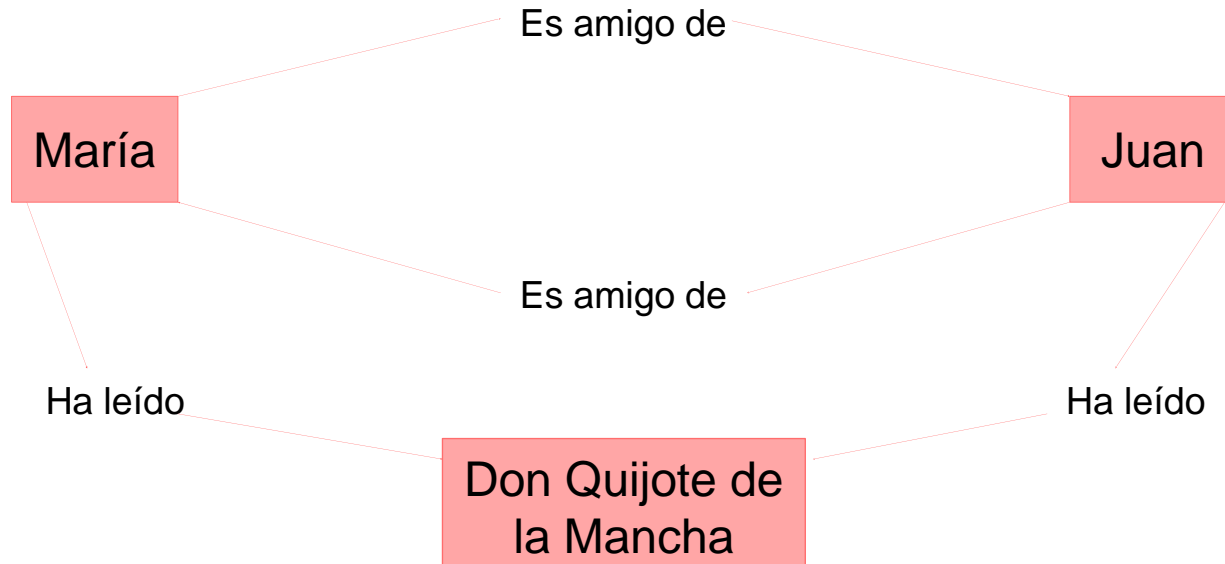
NombreProveedor	ContactoProveedor	Telefono	Producto	Precio
Frutas Gutierrez	Antonio Gutierrez	607454545	Pera	2,05 €
Frutas Gutierrez	Antonio Gutierrez	607454545	Platano	1,87 €
Frutas Gutierrez	Antonio Gutierrez	607454545	Manzana	2,14 €
Frutas Gutierrez	Antonio Gutierrez	607454445	Naranja	1,36 €
Frutas Gutierrez	Antonio Gutierrez	607454545	Limon	1,04 €
Hortalizas del Sur	Guillermo Morales	652854874	Pimiento	0,50 €
Hortalizas del Sur	Guillermo Morales	652854874	Calabaza	1,30 €
Hortalizas del Sur	Guillermo Morales	652884874	Naranja	1,50 €
Hortalizas del Sur	Guillermo Morales	652854874	Pera	3,50 €
Hortalizas del Sur	Guillermo Morales	652854874	Aguacate	5,00 €
Azucarera Sevillana	Rodrigo Mendez	622525885	Azucar(kg)	3,00 €
Azucarera Sevillana	Rodrigo Mendez	622525885	Limon	1,00 €
Azucarera Sevillana	Rodrigo Mendez	622525885	Naranja	1,70 €
Huevos La Puebla	Araceli Arnedo	652879137	Huevo	0,10 €
Huevos La Puebla	Araceli Arnedo	652879137	Pera	1,98 €
Huevos La Puebla	Araceli Arnedo	652879137	Naranja	1,30 €
Huevos La Puebla	Araceli Arnedo	652879137	Arroz (Kg)	1,20 €
Arroces La Cigala	Maria Alvarez	677889922	Arroz (Kg)	1,45 €
Arroces La Cigala	Maria Alvarez	677889922	Naranja	2,45 €
Arroces La Cigala	Maria Alvarez	677889924	Limon	1,30 €



Base de datos orientada a columna

Datos de Proveedor			Datos de Producto	
Nombre	Contacto	Telefono	Nombre	Precio
Frutas Gutierrez	Antonio Gutierrez	607454545	Pera	2,05 €
Frutas Gutierrez	Antonio Gutierrez	607454545	Platano	1,87 €
Frutas Gutierrez	Antonio Gutierrez	607454545	Manzana	2,14 €
Frutas Gutierrez	Antonio Gutierrez	607454445	Naranja	1,36 €
Frutas Gutierrez	Antonio Gutierrez	607454545	Limon	1,04 €
Hortalizas del Sur	Guillermo Morales	652854874	Pimiento	0,50 €
Hortalizas del Sur	Guillermo Morales	652854874	Calabaza	1,30 €
Hortalizas del Sur	Guillermo Morales	652884874	Naranja	1,50 €
Hortalizas del Sur	Guillermo Morales	652854874	Pera	3,50 €
Hortalizas del Sur	Guillermo Morales	652854874	Aguacate	5,00 €
Azucarera Sevillana	Rodrigo Mendez	622525885	Azucar(kg)	3,00 €
Azucarera Sevillana	Rodrigo Mendez	622525885	Limon	1,00 €
Azucarera Sevillana	Rodrigo Mendez	622525885	Naranja	1,70 €
Huevos La Puebla	Araceli Arnedo	652879137	Huevo	0,10 €
Huevos La Puebla	Araceli Arnedo	652879137	Pera	1,98 €
Huevos La Puebla	Araceli Arnedo	652879137	Naranja	1,30 €
Huevos La Puebla	Araceli Arnedo	652879137	Arroz (Kg)	1,20 €
Arroces La Cigala	Maria Alvarez	677889922	Arroz (Kg)	1,45 €
Arroces La Cigala	Maria Alvarez	677889922	Naranja	2,45 €
Arroces La Cigala	Maria Alvarez	677889924	Limon	1,30 €

BBDD no relacionales : orientado a grafos



Modelos de almacenamiento de información

- **Data Warehousing tradicional:** Sistemas diseñados para la integración y análisis de grandes volúmenes de datos estructurados provenientes de múltiples fuentes.
- **Data Lakes:** Almacenes de datos que permiten almacenar datos estructurados y no estructurados a gran escala.
- **Data Marts:** Subconjuntos de data warehouses orientados a departamentos específicos para análisis más focalizados.
- **Comparación entre modelos:** Cada modelo tiene sus propias ventajas y desafíos dependiendo de las necesidades de la organización.

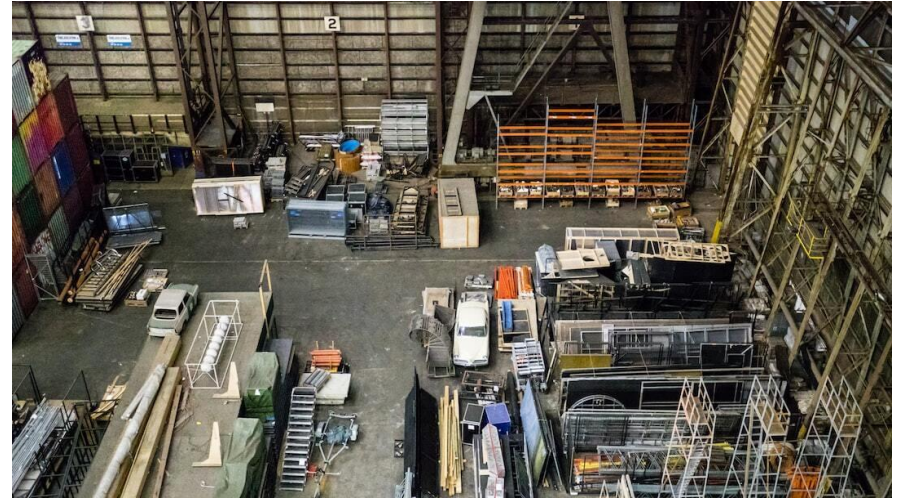


Photo by Charlize on Unsplash

Conceptos fundamentales

- **In Memory:** Almacenamiento de datos directamente en la memoria RAM para acceso rápido y eficiente. Ejemplos incluyen bases de datos como SAP HANA.
- **Online Warehousing:** Almacenamiento de datos en línea, permitiendo el acceso y análisis continuo. Ejemplos incluyen Google BigQuery.
- **Data Virtualization:** Tecnología que permite gestionar datos de diferentes fuentes como si estuvieran en un solo lugar. Ejemplos incluyen Denodo.

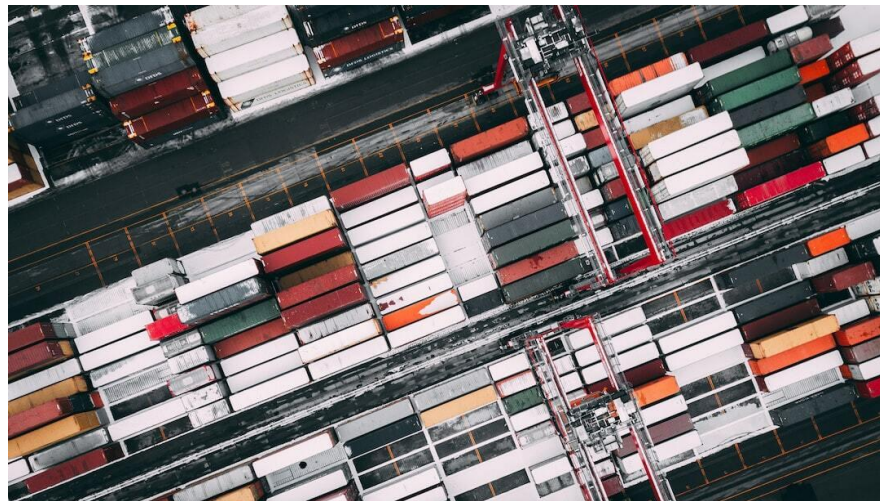


Photo by Erik Odiin on Unsplash

Data Mining

- **Análisis descriptivos:** Métodos para describir y resumir datos históricos, como estadísticas y gráficos.
- **Análisis predictivos:** Técnicas para prever futuros eventos basados en datos históricos y patrones, como modelos de regresión y aprendizaje automático.
- **Análisis prescriptivos:** Estrategias para recomendar acciones basadas en análisis predictivos, como optimización y simulación.

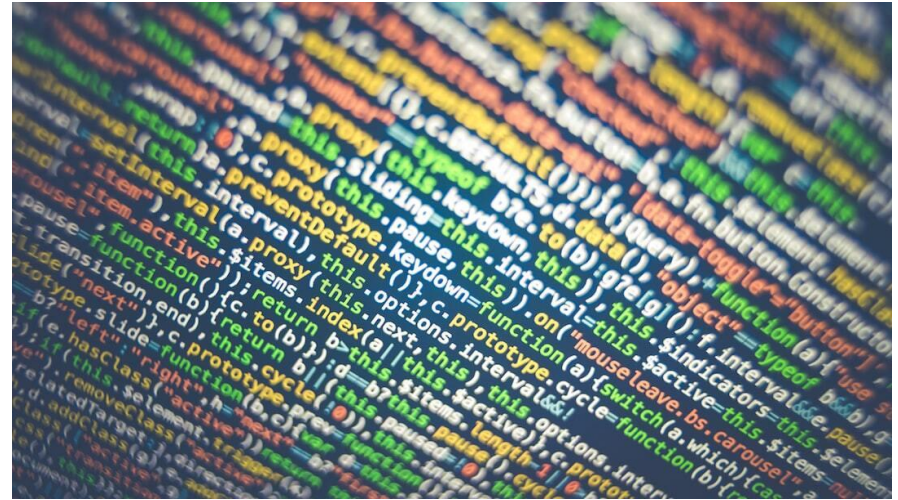


Photo by Markus Spiske on Unsplash

Tendencias tecnológicas

- **Inteligencia Artificial y Machine Learning:** Avances en algoritmos y modelos que permiten análisis y predicciones más precisas y rápidas.
- **Big Data en la nube:** Servicios y plataformas en la nube que facilitan el almacenamiento, procesamiento y análisis de grandes volúmenes de datos.
- **IoT y Big Data:** Integración de datos provenientes de dispositivos IoT para análisis en tiempo real y toma de decisiones informadas.
- **Tecnologías emergentes:** Nuevas herramientas y metodologías que están transformando la forma en que se maneja y analiza el Big Data.

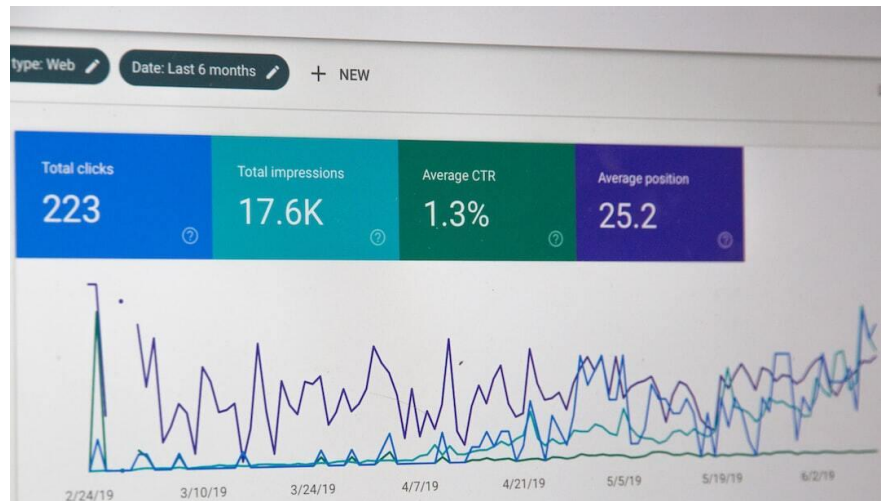


Photo by Stephen Phillips - Hostreviews.co.uk on Unsplash

Visual Discovery

- **Definición e importancia:** Visual Discovery se refiere al proceso de explorar datos a través de representaciones visuales interactivas para descubrir patrones y tendencias.
- **Herramientas y técnicas:** Incluye herramientas como Tableau, Power BI, y técnicas como dashboards interactivos y mapas de calor.
- **Casos de uso:** Aplicaciones prácticas en sectores como salud, finanzas y marketing, donde la visualización de datos ayuda a tomar decisiones informadas.



Photo by Noble Mitchell on Unsplash

Conclusión

- **Resumen de los puntos clave:** Revisión de los temas tratados en la clase, enfatizando la importancia de cada uno en el contexto del Big Data.
- **Importancia del DataWarehouse en el futuro del Big Data:** El DataWarehouse seguirá siendo una pieza clave en la infraestructura de Big Data, adaptándose y evolucionando con nuevas tecnologías.

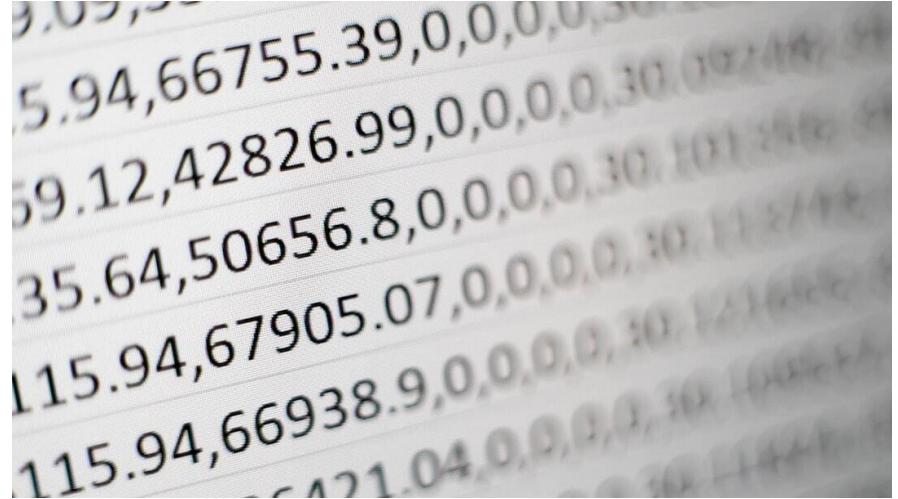


Photo by Mika Baumeister on Unsplash

Ejemplo Práctico: Seguimiento de Vistas de Página

```
CREATE TABLE pageviews (  
  id INT IDENTITY(1,1) PRIMARY KEY,  
  article_id INT NOT NULL,  
  url VARCHAR(255) NOT NULL,  
  views BIGINT DEFAULT 0  
);
```

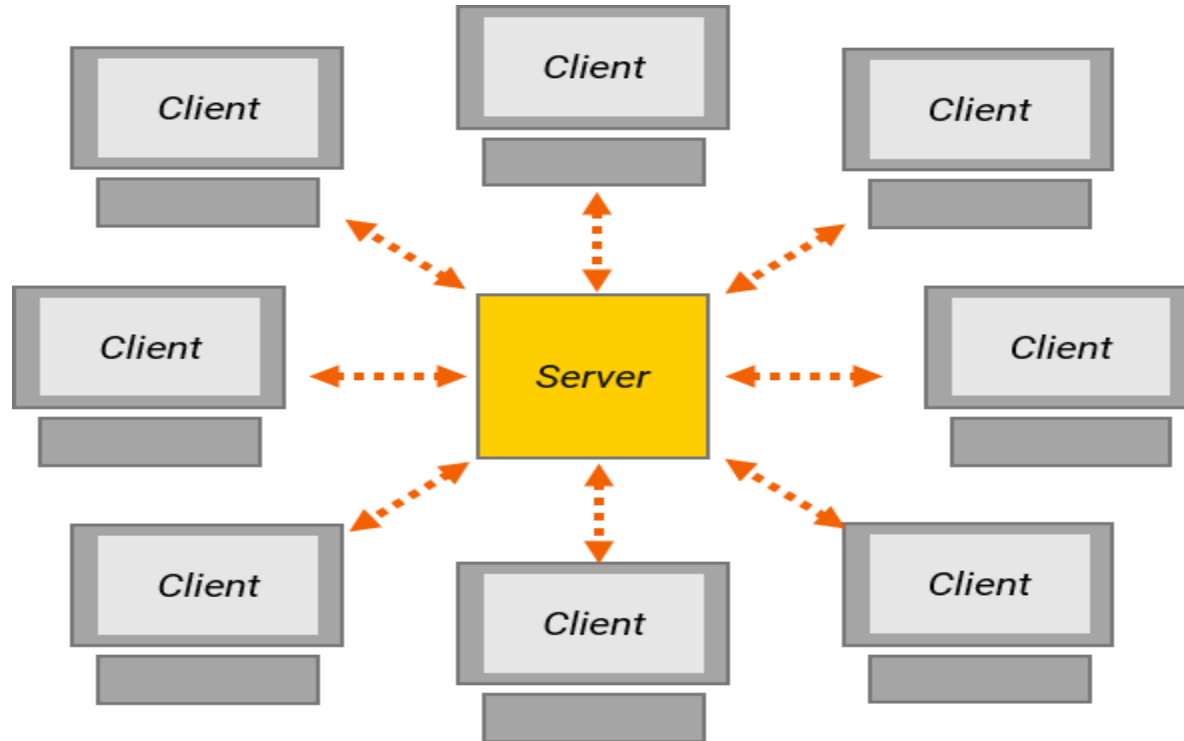
```
UPDATE pageviews SET views = views + 1  
WHERE article_id = 1;
```

PROBLEMAS

- BLOQUEOS
- VELOCIDAD LENTA
- ALMACENAMIENTO

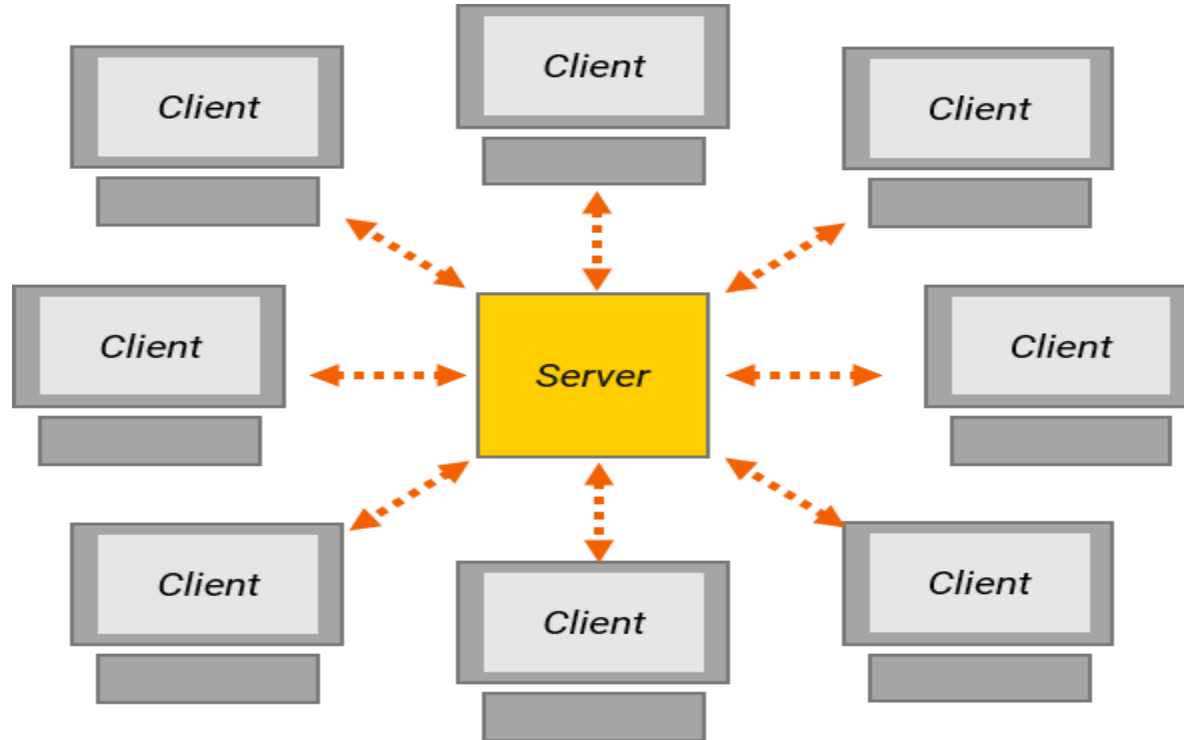
Solución Propuesta: Uso de una Cola

- Estructura de datos: FIFO

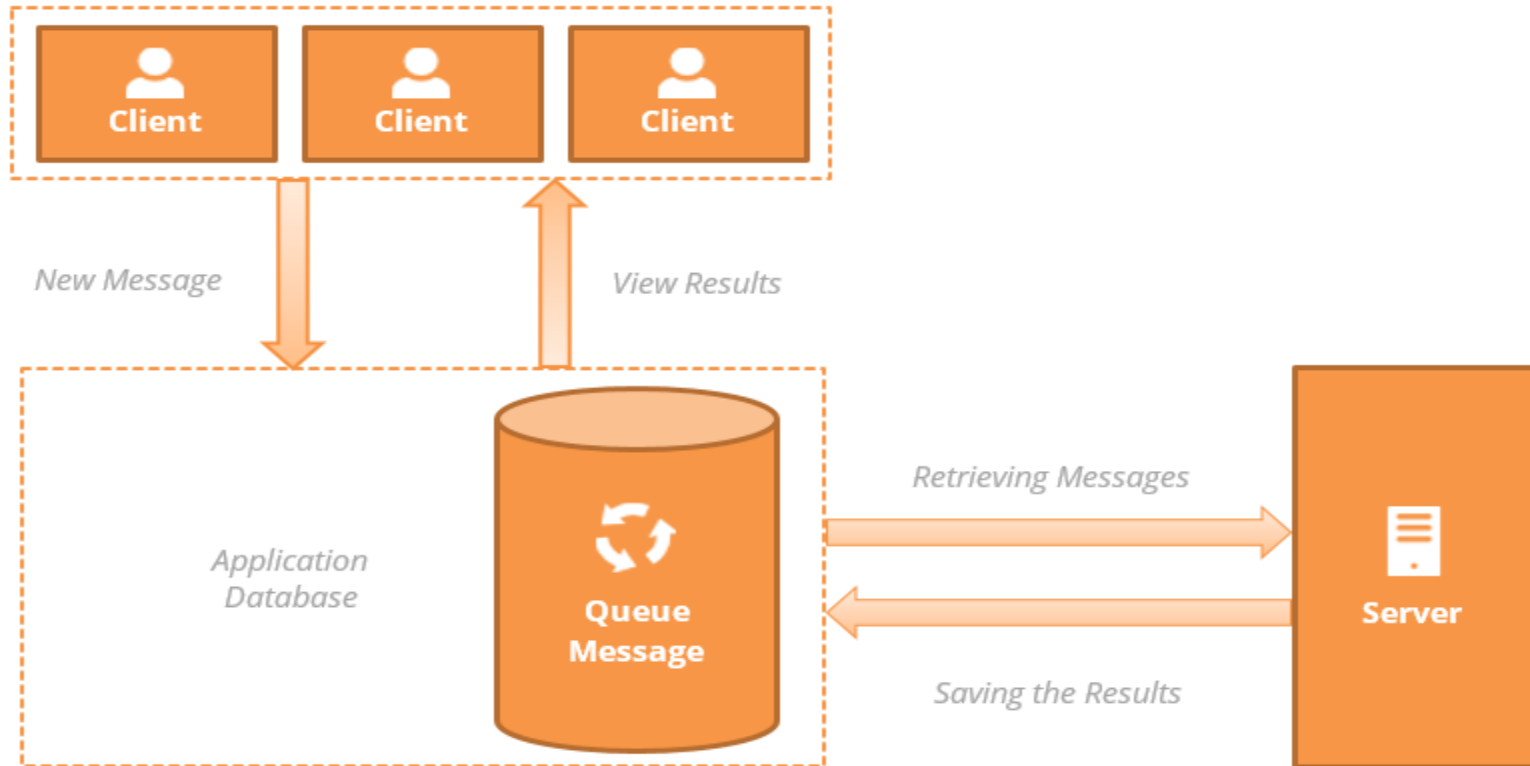


Solución Propuesta: Uso de una Cola

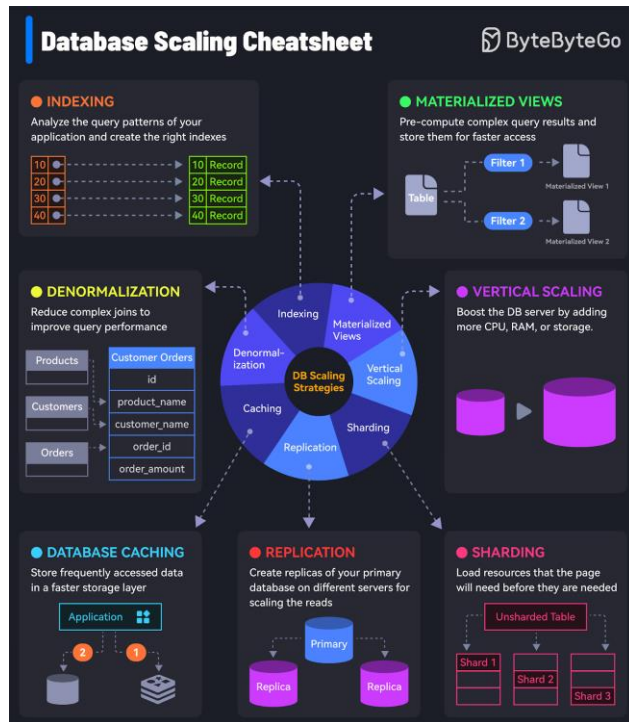
- Estructura de datos: FIFO



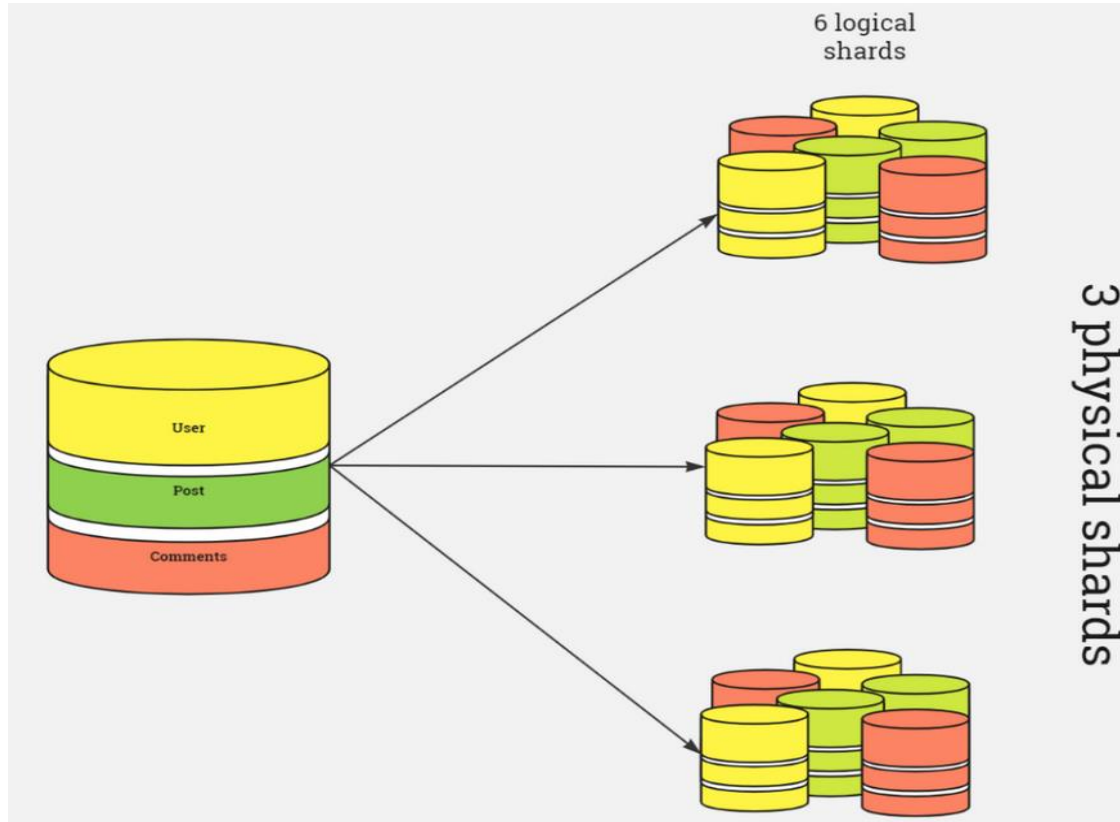
Aplicación : Microservicios



FORMAS DE AUMENTAR LA ESCALA DE UNA BASE DE DATOS



Particiones: lógicos y físicos



Particiones: lógicos y físicos

- .

Customer ID	Name	Age	Fav Food
1	Dolly	11	Egg
2	Leo	8	Beef
3	Pluto	14	Chicken
4	Pickle	10	Egg
5	Juice	12	Pork
6	Major	14	Fish

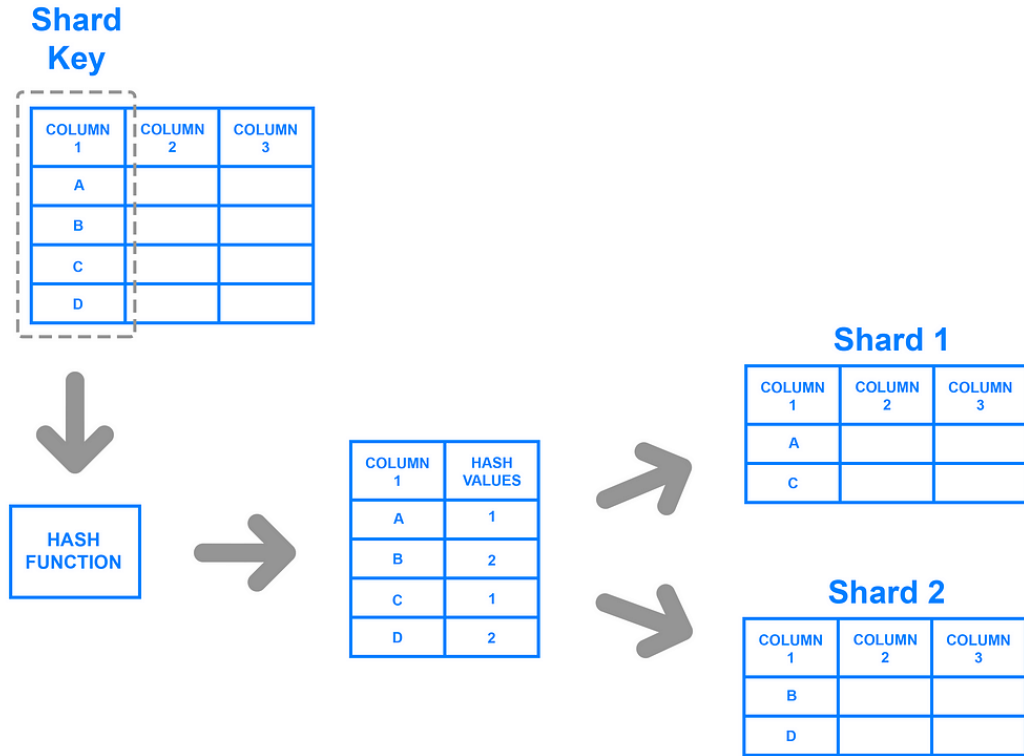
Sharding

•

HP1			
Customer	Name	Age	Fav Food
1	Dolly	11	Egg
2	Leo	8	Beef
3	Pluto	14	Chicken

HP2			
Customer	Name	Age	Fav Food
4	Pickle	10	Egg
5	Juice	12	Pork
6	Major	14	Fish

1) Algorithmico / Hashed Sharding



2) Ranged / Dynamic sharding

PRODUCT	PRICE
WIDGET	\$118
GIZMO	\$88
TRINKET	\$37
THINGAMAJIG	\$18
DOODAD	\$60
TCHOTCHKE	\$999



(\$0-\$49.99)

PRODUCT	PRICE
TRINKET	\$37
THINGAMAJIG	\$18

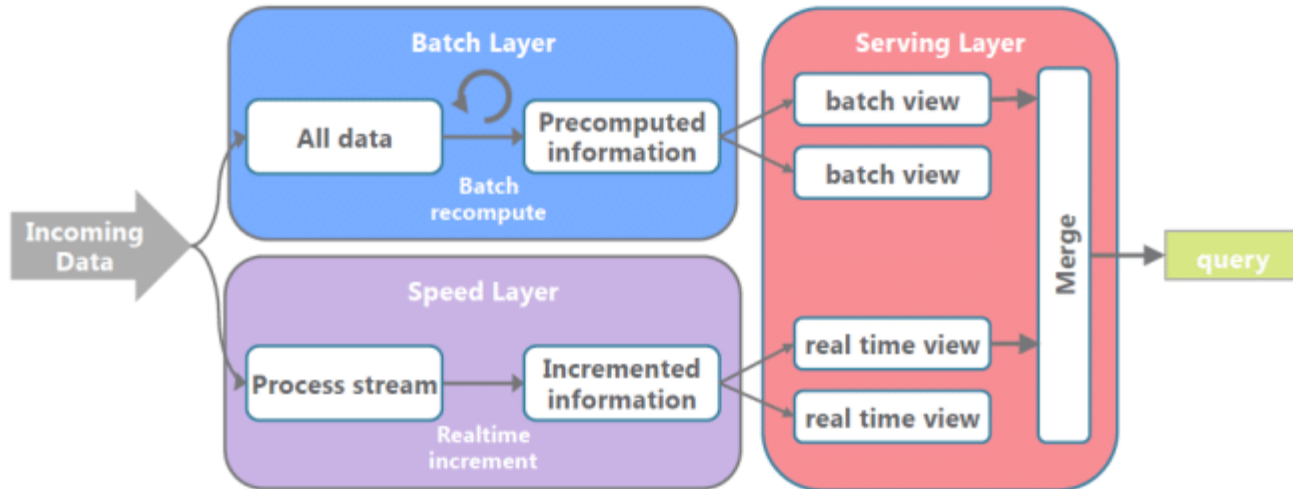
(\$50-\$99.99)

PRODUCT	PRICE
GIZMO	\$88
DOODAD	\$60

(\$100+)

PRODUCT	PRICE
WIDGET	\$118
TCHOTCHKE	\$999

Arquitectura Lambda



Preguntas y discusión

- **Espacio para preguntas:** Tiempo dedicado a resolver dudas y preguntas de los estudiantes sobre los temas tratados en la clase.
- **Discusión abierta:** Fomentar una discusión abierta sobre el futuro del DataWarehouse y su impacto en diferentes industrias.



Photo by SOULSANA on Unsplash