Programa de formación
MACHINE LEARNING AND DATA SCIENCE MLDS

Facultad de
INGENIERÍA

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Módulo 3
Big Data

Unidad 1
Introducción a *Big Data* y *Bases de datos relacionales*

Clase sincrónica

Facultad de
INGENIERÍA

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Bienvenida

Jorge Eliecer Camargo Mendoza, PhD.

UNIVERSIDAD NACIONAL DE COLOMBIA

https://dis.unal.edu.co/~jecamargom/

jecamargom@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial

Facultad de Ingeniería

Universidad Nacional de Colombia

Sede Bogotá

Tabla de contenidos

Objetivos de aprendizaje

Unidad 1 - Introducción a *Big Data*

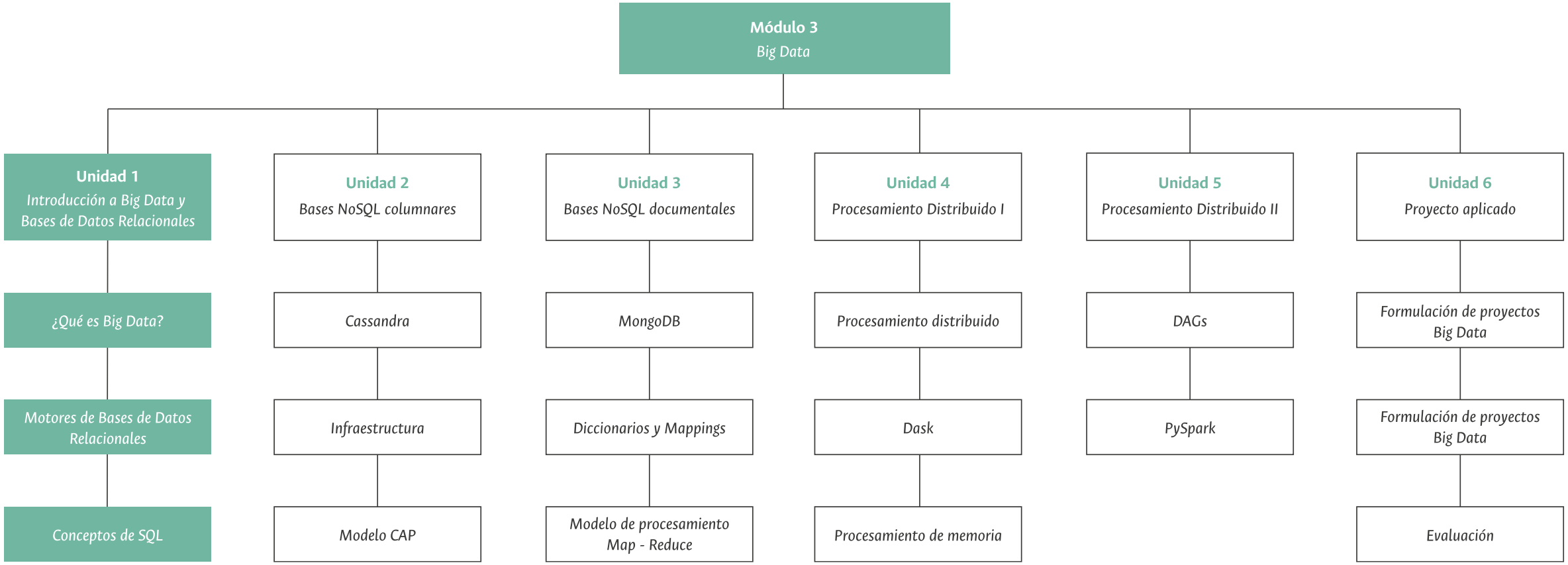
Al finalizar la unidad usted deberá ser capaz de:

Describir de manera precisa los conceptos generales de las distintas herramientas de almacenamiento de grandes cantidades de información y los distintos tipos de bases de datos.

Aplicar operaciones de creación, lectura, actualización y eliminación de datos con el lenguaje de consulta SQL estándar.

Utilizar motores de bases de datos SQLite y PostgreSQL desde Python.

Mapa de contenidos de la unidad



1 Introducción a *Big Data*

Introducción a *Big Data*

Datos

Representaciones simbólicas de atributos o variables que pueden ser cuantitativas o cualitativas.

Temperatura	Valores: 17, 28, 15
Ciudad	Valores: Bogotá, Cartagena
Fecha	Valores: Noviembre 03 2020, 03/11/2020

Introducción a *Big Data*

Datos

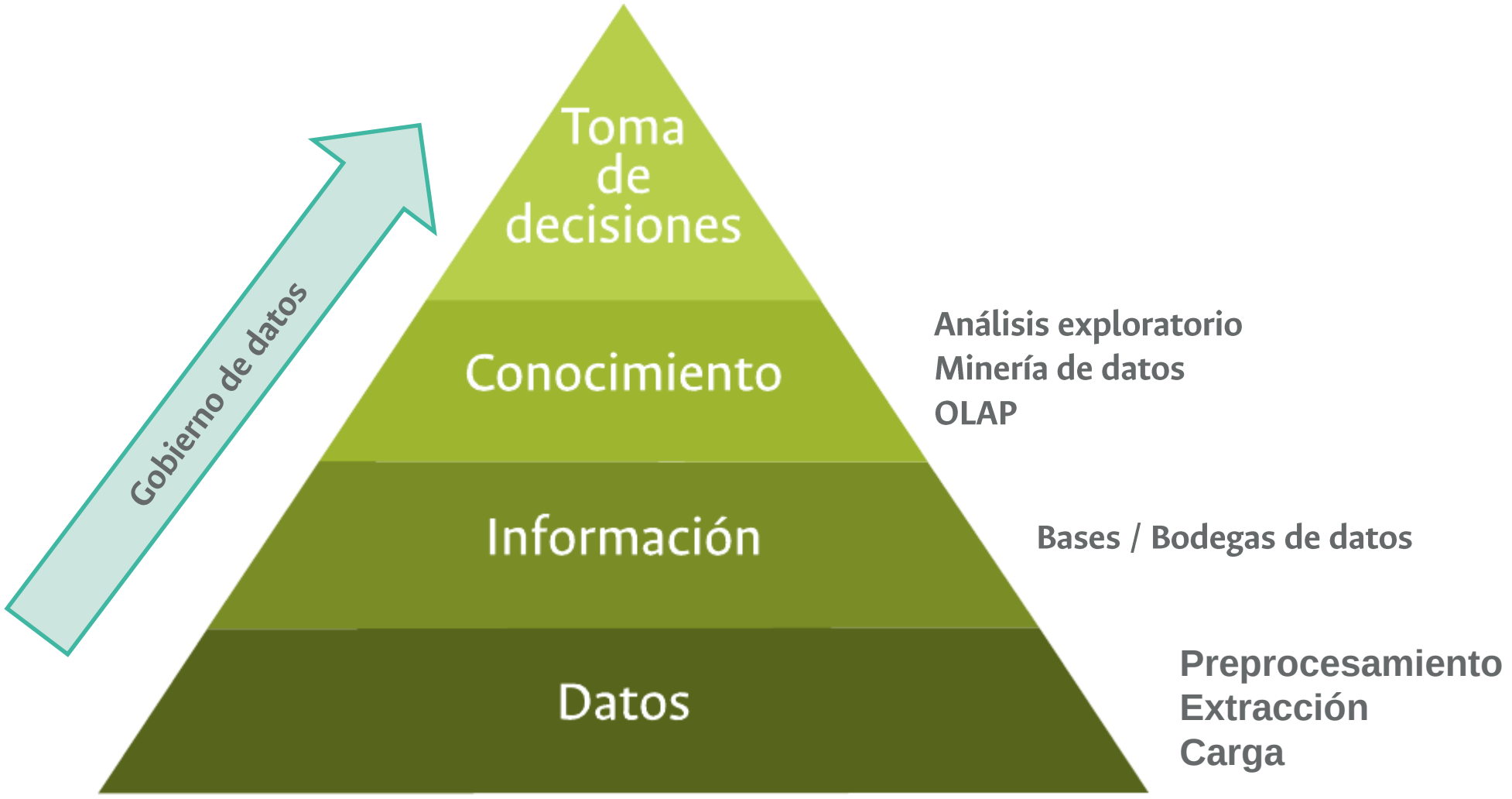
Tradicionalmente almacenados en bases de datos

Datos estructurados y relacionados **"Información"**

2 Jerarquía de los datos

Jerarquía de los datos

Datos



No existe una definición estándar

"***Big Data*** son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos" (Benítez, s.f, p. 11)

" (...) El concepto de ***Big Data*** aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales." (Barranco, 2012, párr. 1)

¿Qué es *Big Data*?

Las 3 V's de *Big Data*

Volumen

Organizaciones masivas, grandes cantidades de datos.

Cada día se tienen tamaños más grandes, actualmente se trabaja en terabytes o incluso *petabytes*.

Cada segundo se tiene un tráfico de internet mayor a **100GB** (2020)

86,000 búsquedas en *Google* cada segundo (2020

Velocidad

Decisiones tardías = Oportunidades perdidas

Funcionamiento en tiempo real, sin almacenar datos. (*Online data analytics*)

Análisis de mercado, seguimiento de pacientes.

Variedad

Múltiples fuentes de distintos tipos de datos.

Procesamiento de datos estructurados o tradicionales (bases de datos) y de datos no / semi estructurados (imágenes, audio, video, ...)

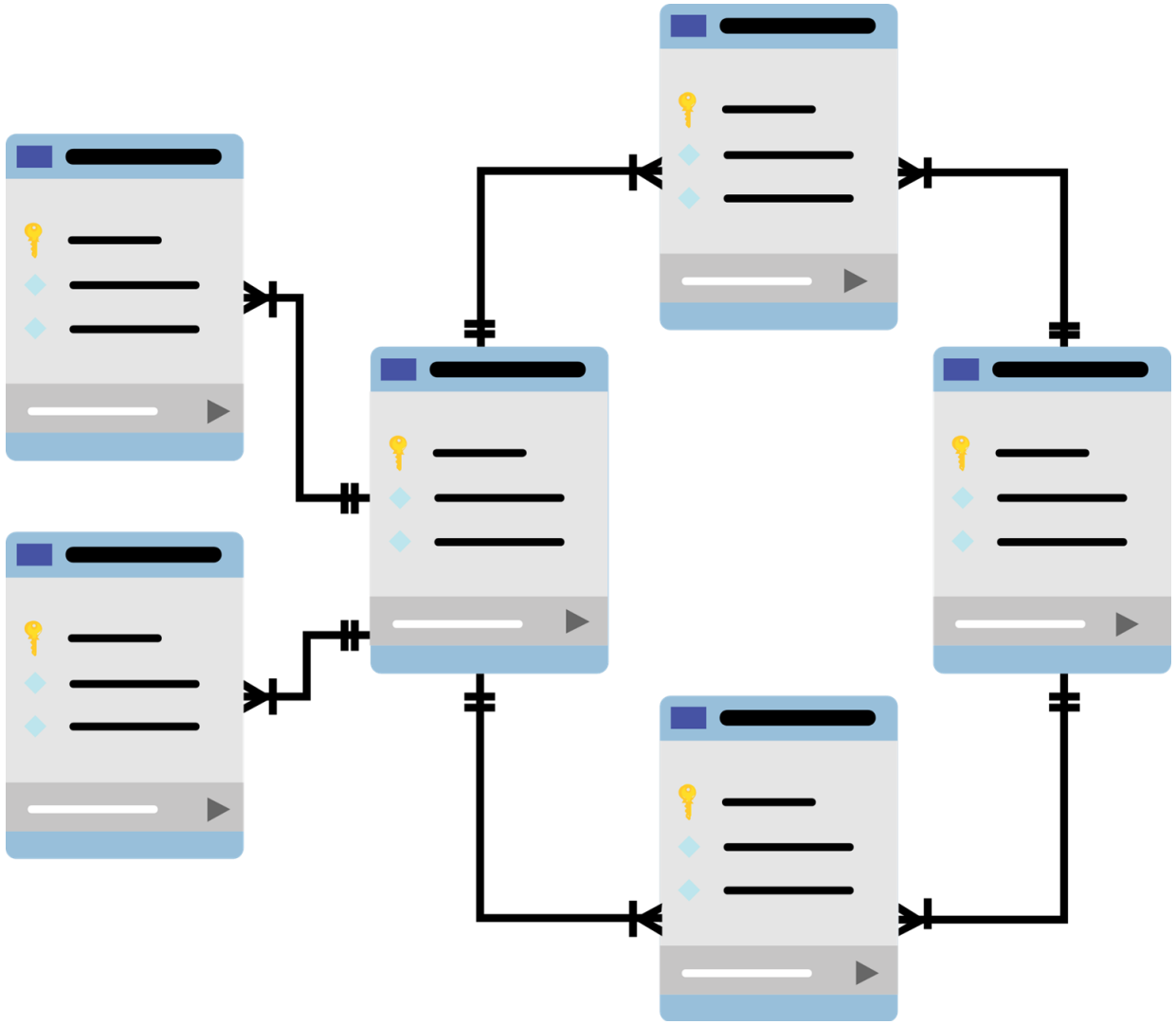
Para datos más actualizados visite: https://www.internetlivestats.com/

3 Aplicaciones

Aplicaciones

- **Segmentación de mercado:**
  - **Rappi** – Separación de usuarios por características, aumento de ventas (focalizadas) y ganancias.
  - **Cadenas de mercado** – Organización establecimientos para el aumento de ventas.
- **Optimización:**
  - **Desempeño físico –** Dispositivos wearable (*Google Fit*, *Apple Watch*).
  - **Rendimiento maquinaria y dispositivos** – Vehículos autónomos, mejoras de redes eléctricas.
  - **Procesos de negocio** – Optimización de stock, Optimización de rutas, Predicción de pérdida de clientes (*Churn*), ...
- **Mercado financiero:**
  - **Compra y venta de acciones** - *High-Frecuency Trading* (HFT).
  - **Puntajes de crédito** - *Credit Scoring.*

4 Conceptos SQL



Conceptos SQL

Bases de datos

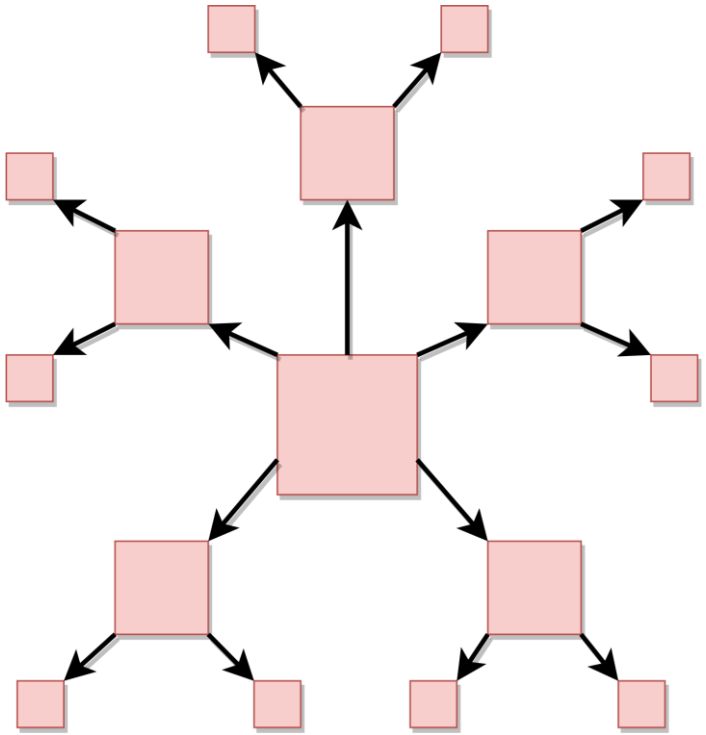
"Una colección de ***datos relacionados***, y una descripción de estos datos, diseñados para cumplir con las necesidades de ***información*** de una organización"
(Connolly & Begg)

Conceptos SQL

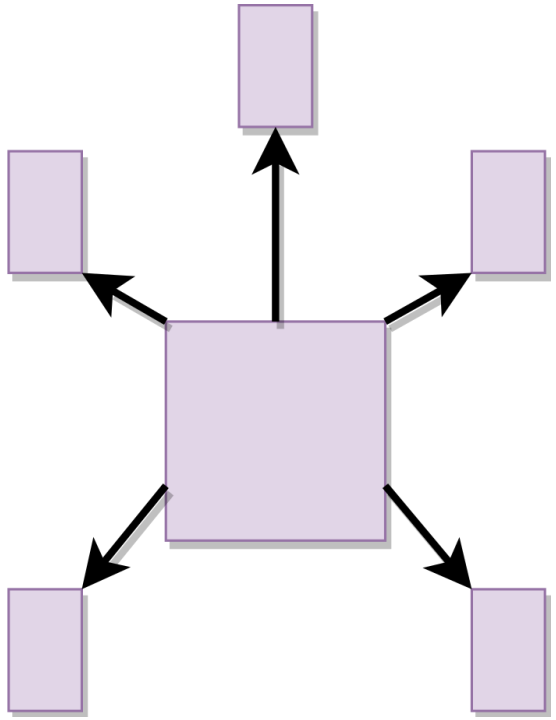
Modelos de *Bases de datos*

1. Jerárquico
2. En red
3. Relacional
4. Entidad-relación
5. Orientada a objetos
6. Documental
7. Copo de nieve (Snowflake)
8. Estrella (Star)
9. Entre otros

Snowflake Schema



Star Schema



Conceptos SQL

Gestor de Bases de datos

"Conjunto de programas que ***maneja*** la ***estructura*** de la BD y ***controla*** el acceso a los datos guardados en ella"

Conceptos SQL

Modelos y *Gestores de Bases de datos*

Tabulares

SQLite

PostgreSQL

MySQL

ORACLE

MariaDB

Microsoft SQL Server

No Tabulares

cassandra

mongoDB.

neo4j

MEMCACHED

amazon DynamoDB

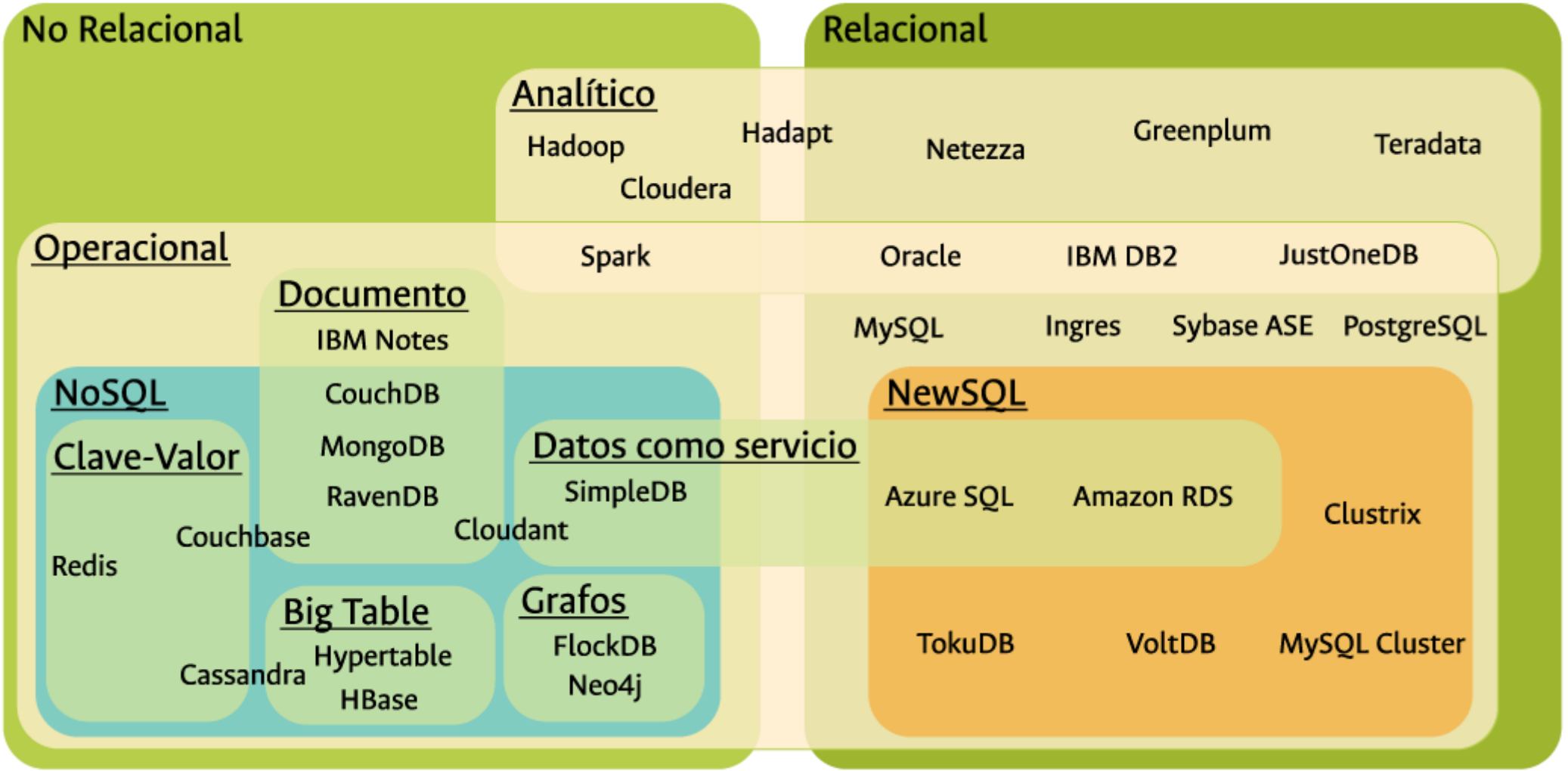
elasticsearch

Google BigQuery

Cloud Firestore

Conceptos SQL

Modelos y *Gestores de Bases de datos*



Almacenamiento de datos (*data storage*)

Data storage			
	Database	**Data warehouse**	**Data lake**
Propósito?	Almacenar, buscar e informar sobre datos estructurados a partir de una única fuente	Almacenar grandes cantidades de datos estructurados de múltiples fuentes en un lugar centralizado	Almacena datos estructurados, semiestructurados y no estructurados, lo que permite almacenar datos crudos de todas las fuentes sin necesidad de procesarlos o transformarlos en ese momento
Cantidad de información procesada en promedio	Relativamente pequeña comparando con Warehouses y Data lakes	Terabytes	Petabytes

Conceptos SQL

Modelo de datos Relacional - SQL

En el modelo de datos relacional, los datos se almacenan de manera estructurada en un esquema o plantilla, la cual comúnmente se representa en tablas.

Estas tablas, típicamente, modelan una entidad del mundo real y representan una relación entre los distintos datos que contienen; también permiten modelar relaciones entre entidades reales, es decir, una tabla se puede relacionar con otras tablas.

Conceptos SQL

¿Qué es SQL?

Structured Query Language más conocido como SQL (por sus siglas en inglés) es un lenguaje de bases de datos que aterriza las operaciones del álgebra relacional, las cuales son utilizadas en los modelos de bases de datos relacionales.

Este lenguaje cuenta con palabras reservadas que son utilizadas para expresar distintas operaciones que permiten trabajar sobre las tablas o entidades.

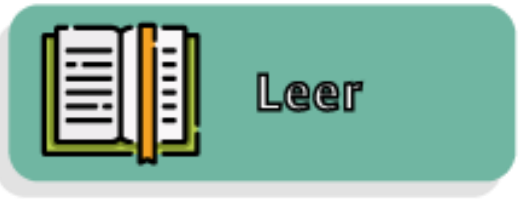
Conceptos SQL

Operaciones CRUD

CREATE


```
INSERT INTO
  libro(lib_anio, lib_nombre)
VALUES
  (1867 ,“Maria”);
```

READ

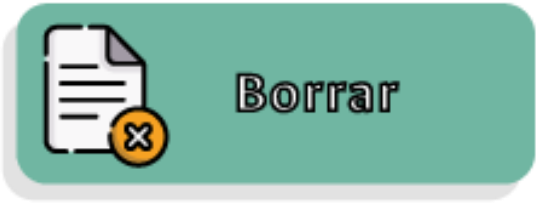



```
SELECT
   lib_anio,
   lib_nombre
FROM
   libro;
```

UPDATE


```
UPDATE
   libro
SET
   anio = 1956
WHERE
   lib_id = 1051;
```

DELETE




```
DELETE
FROM
   libro
WHERE
   lib_id = 156;
```

Conceptos SQL

Operaciones CRUD

Crear

En SQL, con la sentencia *INSERT*, se introduce dentro de una tabla un nuevo registro.

```
CREATE TABLE libro(
  lib_id INT,
  lib_nombre VARCHAR(255),
  lib_anio INT);

INSERT INTO
  libro(lib_id,lib_nombre,lib_anio)
VALUES
  (100051,'Cien anios de soledad',1956);
```

lib_id	lib_nombre	lib_anio
100051	Cien anios de soledad	1967
200032	La voragine	1924
300033	Maria	1867
401156	Condores no entierran todos los dias	1971

Conceptos SQL

Operaciones CRUD

Leer

En SQL, con la sentencia *SELECT*, se hace referencia a la recuperación de datos o conjuntos de registros de una o varias tablas.

```
SELECT
   lib_nombre
FROM
   libro
```

lib_id	lib_nombre	lib_anio
100051	Cien anios de soledad	1967
200032	La voragine	1924
300033	Maria	1867
401156	Condores no entierran todos los dias	1971

Conceptos SQL

Operaciones CRUD

En SQL, con la sentencia *UPDATE*, se modifican los datos de un registro específico dentro de una tabla:

```
UPDATE
  libro
SET
  lib_anio = 1956
WHERE
  lib_id = 100051;
```

lib_id	lib_nombre	lib_anio
100051	Cien anios de soledad	1956
200032	La voragine	1924
300033	Maria	1867
401156	Condores no entierran todos los dias	1971

Conceptos SQL

Operaciones CRUD

Borrar

En SQL, con la sentencia *DELETE*, se eliminan registros específicos dentro de una tabla:

```
DELETE
FROM
  libro
WHERE
  lib_id = 401156;
```

lib_id	lib_nombre	lib_anio
100051	Cien anios de soledad	1967
200032	La voragine	1924
300033	Maria	1867

Conceptos SQL

Principio ACID

Atomicidad

(*Atomicity*) Todos los cambios sobre los datos (transacciones) son realizados como uno solo; es decir, todos los cambios de la transacción se realizan o ninguno lo hace.

Consistencia

(*Consistency*) Los datos son consistentes al inicio y al final de una transacción.

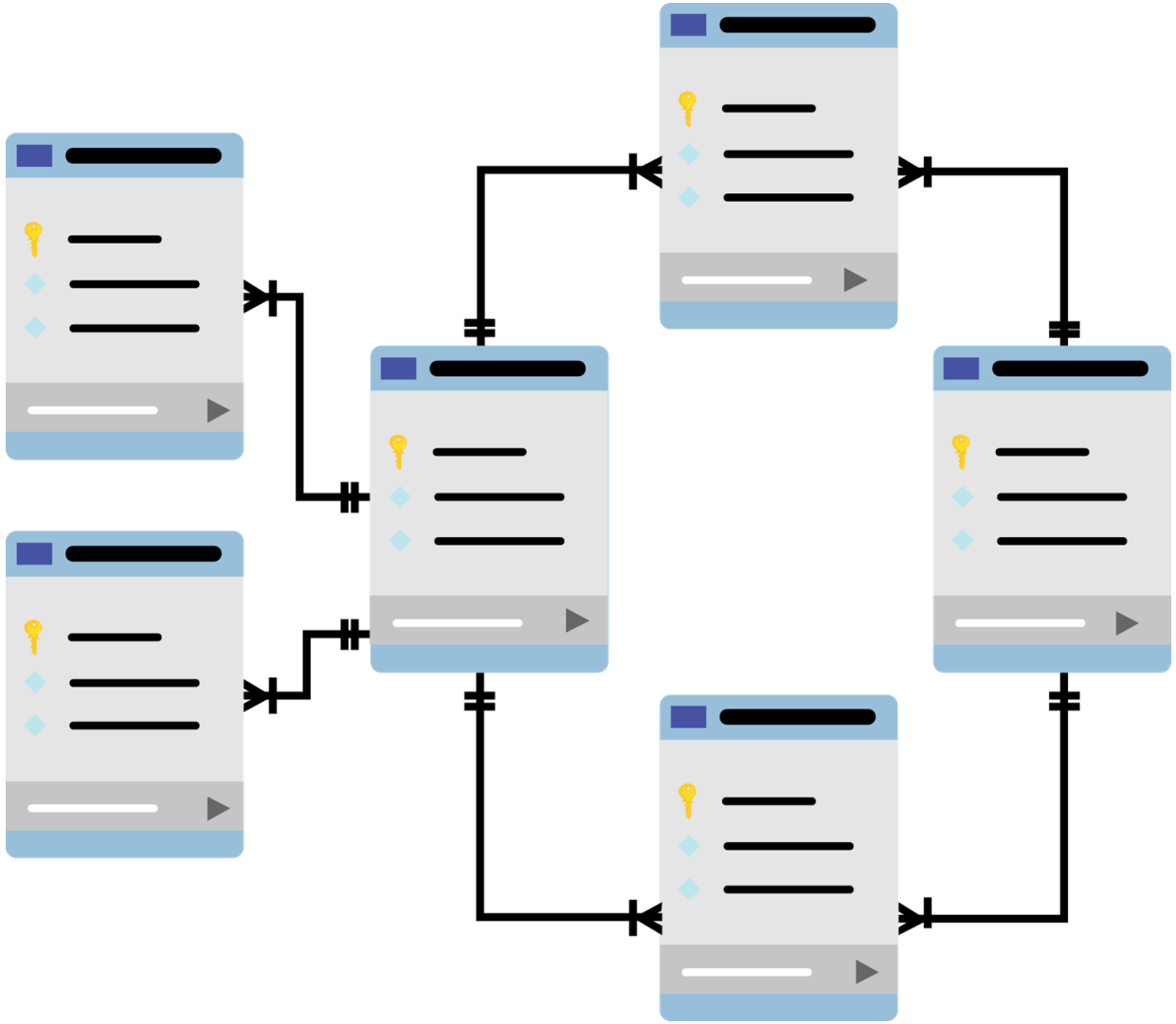
Aislamiento

(*Isolation*) El estado intermedio de una transacción es invisible para otras transacciones.

Durabilidad

(*Durability*) Después de que una transacción se completa satisfactoriamente, los cambios sobre los datos son persistente y no se pueden deshacer.

5 Actividades (SQL)



- **Talleres guiados**:
  - SQLite 3
  - PostgreSQL
  - Conceptos de SQL
  - Integración Pandas y SQL

- **Taller SQL**:
  - Consultas con SQL. (*UNCode*)

Despedida

¡Gracias por su atención!

Jorge Eliecer Camargo Mendoza, PhD.

UNIVERSIDAD NACIONAL DE COLOMBIA

https://dis.unal.edu.co/~jecamargom/

jecamargom@unal.edu.co

Departamento de Ingeniería de Sistemas e Industrial

Facultad de Ingeniería

Universidad Nacional de Colombia

Sede Bogotá

Referencias

Barranco, (2012), ¿Qué es Big Data?, https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html

Benítez J., (s. f.), Big Data: Algoritmos tecnología y aplicaciones, http://madm.uib.es/wp-content/uploads/2016/06/Jose-Manuel-Benitez-Sanchez-Big-Data-Algoritmos_tecnologia_y_aplicaciones.pdf

Ladrero I., (2018), 10 ejemplos de usos reales de Big Data Analytics, https://www.baoss.es/10-ejemplos-usos-reales-big-data/

Kashmir H., (2012), How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did, https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#5afd37d46668

Contxto, (2020), ¿Qué hacen las apps de entrega, como Rappi, para ganar en América Latina?, https://contxto.com/es/colombia-es/app-entregas-rappi-exito-america-latina/

Derechos de imágenes

Macrovector (s.f) Vector de Textura https://www.freepik.es/vectores/textura

8photo (s.f) Foto de Personas https://www.freepik.es/fotos/personas'

Fabian Krüger https://pixabay.com/es/users/kruegerfotografie-6309059/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=3178765

Macrovector. (s.f.). Gran colección de elementos de procesamiento de datos vector gratuito. [Vector]. https://www.freepik.es/vector-gratis/gran-coleccion-elementos-procesamiento-datos_7439564.htm#page=1&query=servidores&position=49#&position=49

Vectorpouch. (s.f.). Página de inicio isométrica de noticias mundiales. Planeta terrestre en una enorme pantalla de smartphone con presentadores de televisión emitiendo en televisión. [Vector]. https://www.shutterstock.com/es/image-vector/world-news-isometric-landing-page-earth-1566274144

Derechos de imágenes

Vecteezy. (s.f.). Concepto del mercado de valores. [Vector]. https://es.vecteezy.com/arte-vectorial/478935-concepto-del-mercado-de-valores

Pikisuperstar. (2020). Image upload concept for landing page Free Vector. [Vector]. https://www.freepik.com/free-vector/image-upload-concept-landing-page_5566770.htm

Freepik. (s.f.). Ilustración del concepto de sistema operativo. [Vector]. https://www.freepik.es/vector-gratis/ilustracion-concepto-sistema-operativo_7967803.htm

Freepik. (s.f.). Coding free icon. [Icono]. https://www.flaticon.com/free-icon/coding_1969984

Google Cloud Datastore Logo. [Icono]. Google Cloud Offficial Icons & Solution Architectures. https://cloud.google.com/icons

Créditos

Facultad de
INGENIERÍA

Autores
Jorge Eliecer Camargo Mendoza, PhD

Asistente docente
Leonardo Avendaño Rocha
Alberto Nicolai Romero Martínez
Rosa Alejandra Superlano Esquivel
Edder Hernández Forero
Brian Chaparro Cetina

Diseño instruccional
Claudia Patricia Rodríguez Sánchez

Diseño gráfico
Clara Valeria Suárez Caballero
Milton R. Pachón Pinzón

Diagramadora PPT
Daniela Duque

Fecha
2022-II