

# Banco de Dados NoSQL



Prof. Dr. Vladimir Costa Alencar

LANA - UEPB

<https://www.valencar.com/>

@vladimiralencar (twitter)

Fevereiro, 2019



LANA  
DATA ANALYSIS LABORATORY



# Data Science e Big Data



## Tipos de Dados: Estruturados

**Tabela**

Veiculo

**Atributos**

Placa	Fabricante	Marca	Ano	Cor
IOS-0078	Renault	Sandero	2009	Vermelho
ITO-1314	Volkswagen	Fox	2010	Azul
IJM-1453	Hyundai	I30	2014	Pérola
IVA-2018	Chevrolet	Onix	2015	Branco
MAI-1852	Citroen	C3	2013	Preto

**Tuplas**

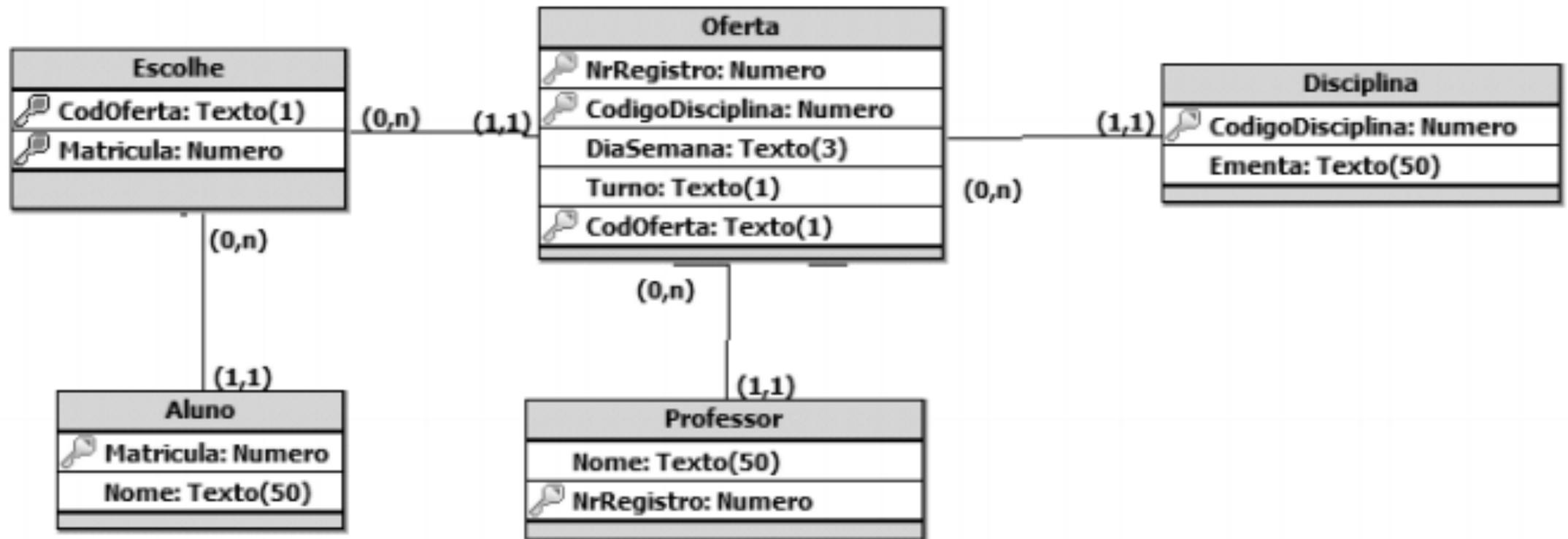
**Domínio**

A diagram illustrating a table structure. The word 'Atributos' is positioned above the table, with red arrows pointing to each of the five columns: Placa, Fabricante, Marca, Ano, and Cor. The word 'Tuplas' is positioned to the left of the table, with red arrows pointing to each of the five rows. A red arrow points from the word 'Domínio' at the bottom to the 'Ano' column.

# Data Science e Big Data



## Tipos de Dados: Estruturados – Banco de Dados Relacionais



# Data Science e Big Data



## Tipos de Dados: Semi-Estruturados - XML

```
<recipe id="117" category="sobremesa">  
<title> Bolo de banana </title>  
<author>  
<email> Miguel Furtado furtado@predialnet.com.br </email>  
</author>  
<date>Sunday,04 Jun 2000</date>  
<description> O bolo de banana é feito com banana prata e  
possui um sabor maravilhoso.  
Pode ser servido quente ou frio. </description>  
<ingredients> ... </ingredients>  
<preparation>  
Combine tudo no liquidificador e bata até misturar bem.  
Leve ao forno por 20 minutos. Pronto. É só servir! </preparation>  
<related url="#BoloChocolate">Bolo de Chocolate</related></recipe>
```



# Semi-Estruturado – Arquivo .JSON (Java Script Object Notation)

```
{
  "id_str": "694877215301873666",
  "text": "When Ebola performs duty in Sierra Leone,US UK "Warships" were sent 2 treat patients. Now Zica: sending 2 #Brazil???",
  "url": "https://t.co/wTW61kKsI0",
  "id": 694877215301873666,
  "source": "<a href='\"http://twitter.com\"' rel='\"nofollow\"'>Twitter Web Client</a>",
  "coordinates": 10000,20000,
  "retweet_count": 0,
  "followers_count": 722,
  "description": "Portfolio Manager BehindtheScenes Money Maker Management+",
  "friends_count": 165,
  "location": "ATHENS",
  "screen_name": "anthonysamaha",
  "lang": "en",
  "favourites_count": 4963,
  "name": "Anthony Samaha",
  "url": "http://t.co/o6uil766Ds",
  "time_zone": "Athens",
  "lang": "en",
  "created_at": "Wed Feb 03 13:36:52 +0000 2016",
  "place": Athens,
  "iso_language_code": "en"
}
```





# Banco de Dados Relacionais

Armazena dados em **tabelas**

Possuem atributo tipo **chave**

implementam funcionalidades simples do tipo **CRUD** (do inglês *Create, Read, Update e Delete*)

controla o armazenamento, recuperação, exclusão, segurança e integridade dos dados

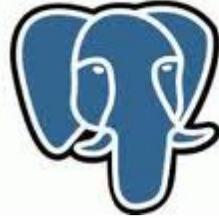
Tabela Empregado

cod_empr	nome	cod_depto
1	paula	1
2	jaque	1
3	andre	2
4	luiz	2

# Os SGBD Relacionais



PostgreSQL





# Banco de Dados Relacionais

Armazena dados em **tabelas**

Possuem atributo tipo **chave**

implementam funcionalidades simples do tipo **CRUD** (do inglês *Create, Read, Update e Delete*)

controla o armazenamento, recuperação, exclusão, segurança e integridade dos dados

Tabela Empregado

cod_empr	nome	cod_depto
1	paula	1
2	jaque	1
3	andre	2
4	luiz	2

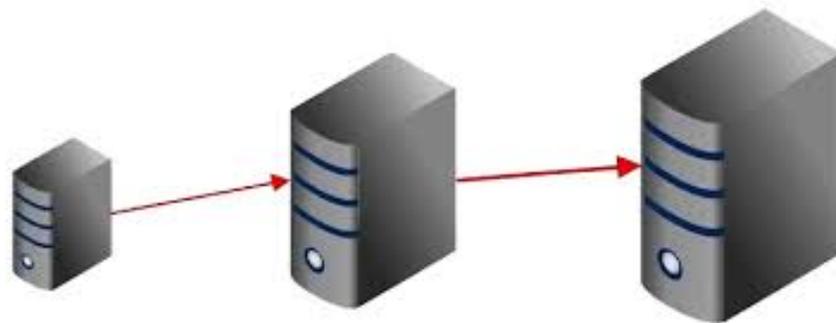


# Banco de Dados Não-Relacionais (NoSQL) Escalabilidade

---

Volume de dados crescente

Disponibilidade dos dados (tempo real)

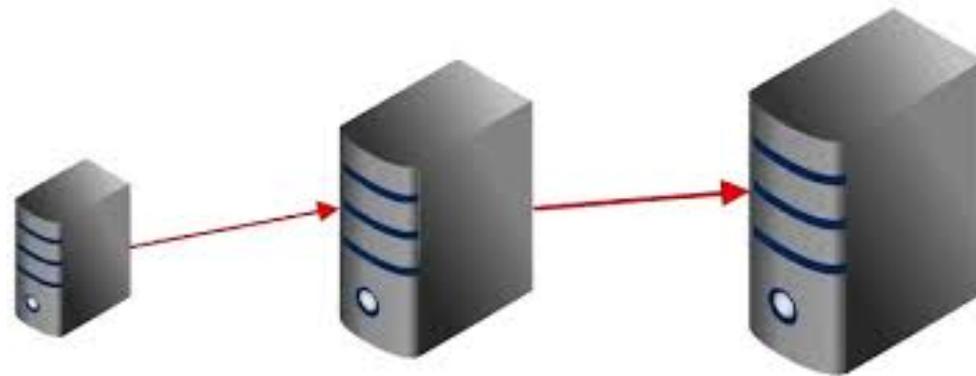


# Escalabilidade

---

Característica de **Aumentar a capacidade** do sistema (seja processamento, armazenamento, E/S)

Há um Aumento da **Performance** depois da adição de **Hardware ou Serviço**.



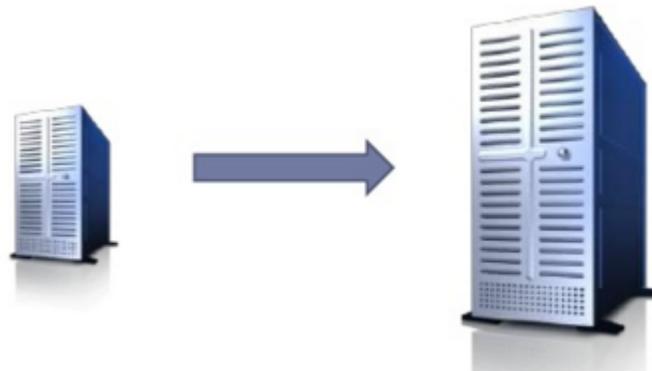
# Escalabilidade Vertical

---

Adicionar novos componentes na máquina.

É quando você coloca mais memória, mais disco, mais CPU no seu servidor.

Geralmente requer desligar a máquina, adicionar recursos e ligar novamente.



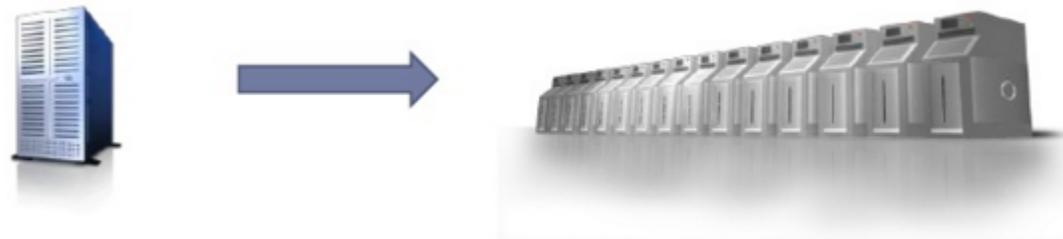
# Escalabilidade Horizontal

---

Adicionar máquinas em Paralelo

Coloca mais servidores para atender a demanda

Carga é balanceada entre os servidores  
(ex. Cluster)



# NoSQL- Escalabilidade

---



**Escalabilidade Vertical  
(Scale Up)**



**Escalabilidade Horizontal  
(Scale Out)**

# NoSQL – Não Relacionais

---

São Facilmente escaláveis Horizontalmente

Trabalham com quantidades maciças de dados (Big Data)

Não utilizam o modelo relacional para suas estruturas de dados

NoSql – Not Only Sql





# Banco de Dados Não-Relacionais (NoSQL)

---

Gerenciar os grandes volumes de dados

Buscar alto desempenho e disponibilidade

Permitem uma escalabilidade mais barata e menos trabalhosa

Características de poder trabalhar com dados semi-estruturados ou crus vindos de diversas origens

(arquivos de log, web-sites, arquivos multimídia, JSON, etc...)

Modelos Baseados em Documentos, Colunas, Chave-Valor e Grafos.



# Tipos (NoSQL)

---

**Chave/Valor (Key/Value)**

**Orientados a Documentos**

**Orientados a Colunas**

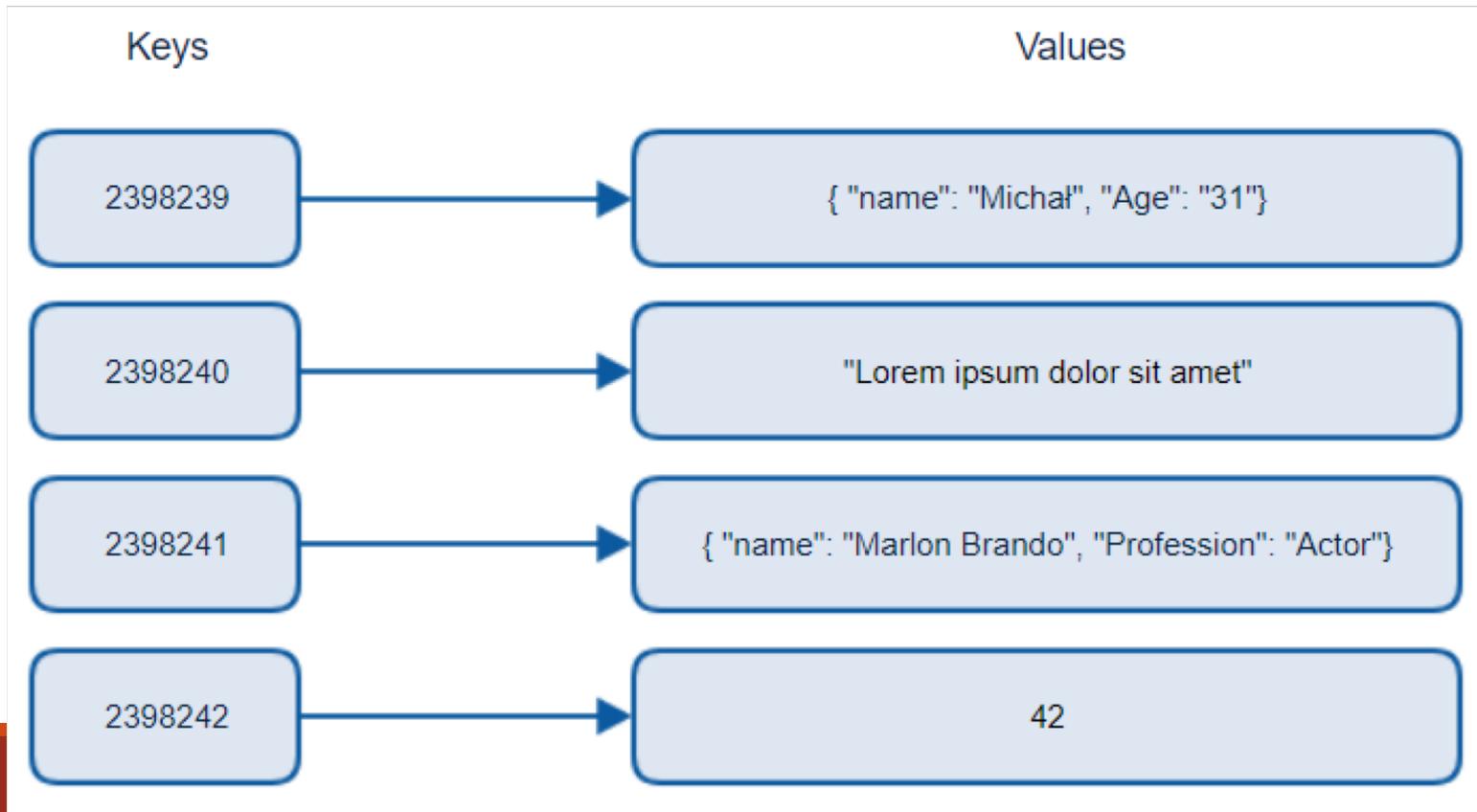
**Grafos**

# NoSQL – Chave/Valor

O modelo chave-valor é o mais simples e fácil de implementar

---

Usa uma tabela hash na qual há uma chave única e um indicador de um dado ou de um item em particular



# NoSQL – Chave/Valor - **Aplicações:**

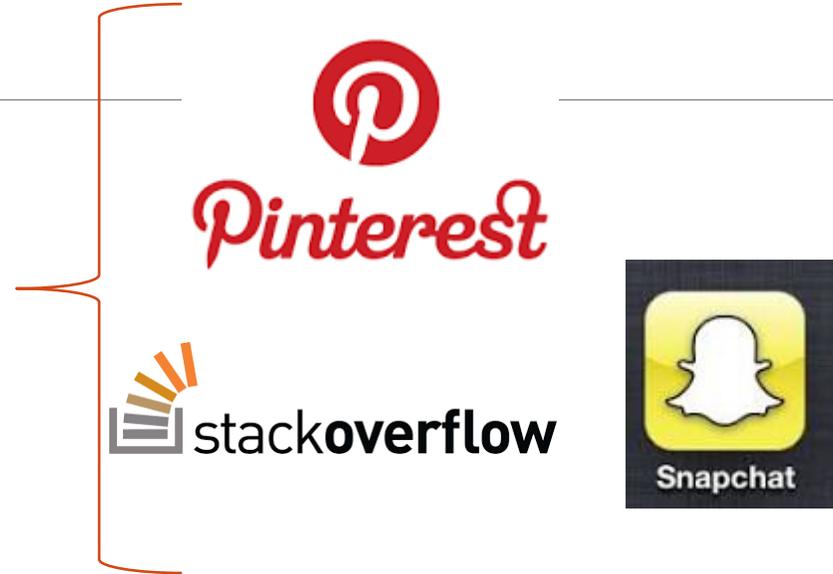
---

Cache de conteúdo  
(grande quantidade de dados, carregamento massivo)

Pesquisas rápidas

Logging  
(registro de eventos relevantes)

# NoSQL – Chave/Valor - uso:



## NoSQL – Orientados a Colunas

---

Enquanto um banco de dados relacional é otimizado para armazenar **linhas** de dados, geralmente para aplicativos transacionais, um banco de dados colunar é otimizado para **recuperação rápida de colunas** de dados, normalmente em aplicativos analíticos

Armazena e processa grandes quantidades de dados distribuídos em muitas máquinas.

Reduz expressivamente os requisitos gerais de E/S de disco e diminui a quantidade de dados que você precisa carregar do disco.

### Original data

oid	pid	cust	date	price
1	12	Sam	1/1/06	\$100
2	17	Mike	3/4/06	\$87
3	18	Joe	1/2/06	\$12
4	4	Andy	8/4/06	\$125

sales

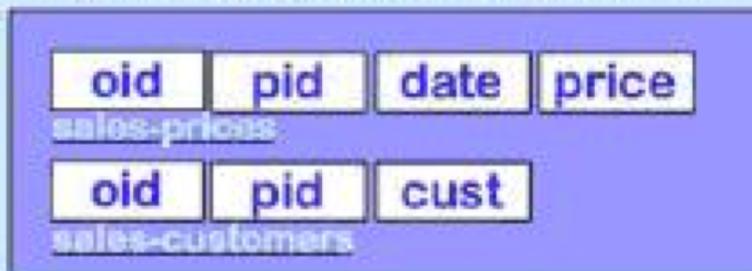


### Physically Stored as Columns

oid	pid	cust	date	price
1	12	Sam	1/1/06	\$100
2	17	Mike	3/4/06	\$87
3	18	Joe	1/2/06	\$12
4	4	Andy	8/4/06	\$125



### Split into Several Projections



# NoSQL – Orientados a Columnas

# NoSQL – Orientados a Colunas- Aplicação

---

Aplicações distribuídas com uso intensivo de dados

Ex. Facebook



# NoSQL – Orientados a Columnas - Uso



# NoSQL – Orientados a Documentos

---

Um documento é:

- um objeto

- tem um identificador único

- tem um conjunto de campos, que podem ser strings, listas ou documentos aninhados

Em cada documento temos um conjunto de campos (chaves) e o valor deste campo

# NoSQL – Orientados a Documentos

---

## Documento: archivo .JSON

```
{ "screen name": "@infomagazine"  
  "text": "The 8 most stressful jobs in tech. Is yours one of them?"  
  "url": "https://t.co/qOfPJJrfcw",  
  "screen name": @infomagazine  
  "lang": "En",  
  "created_at": "Tue Feb 05 02:31:38 +0000 2019",  
  "place": "United States", ...  
}
```

# NoSQL – Orientados a Documentos – Aplicação

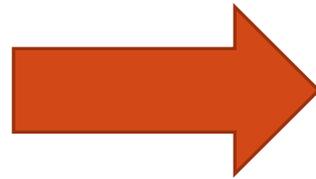
---

Aplicações Web

(Similar ao armazenamento chave-valor)

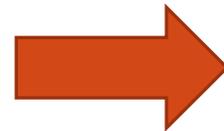
Tolerante a dados incompletos

# NoSQL – Orientados a Documentos - Uso



Linked in

Mc  
Graw  
Hill  
Education



Expedia®

ebay®

NOKIA  
Connecting People

foursquare®

# NoSQL – Grafos

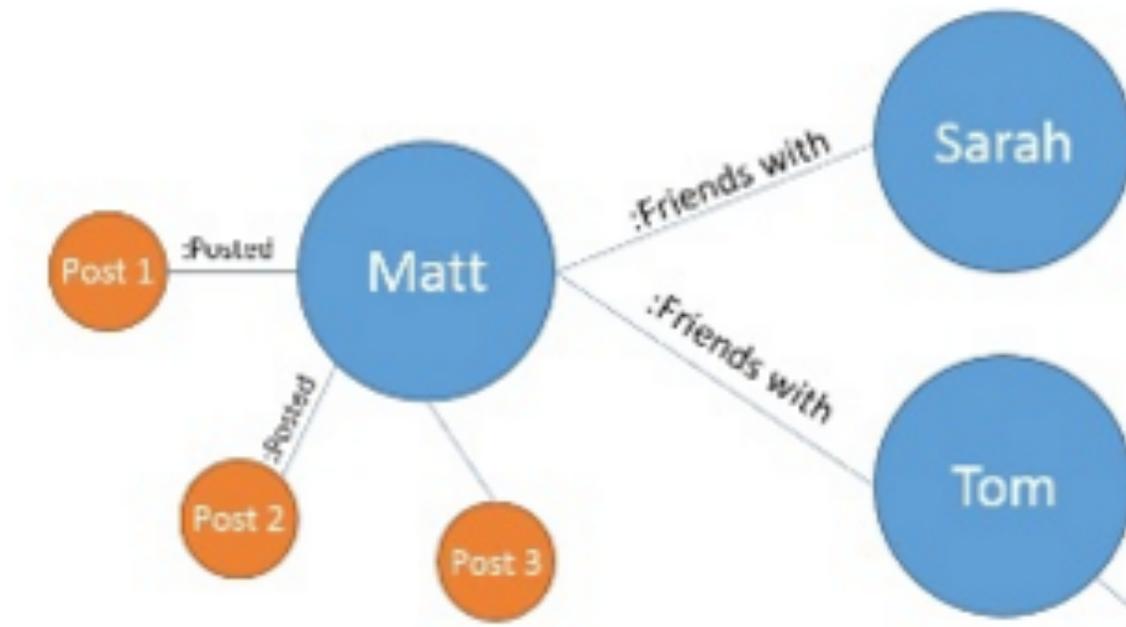
Possuem:

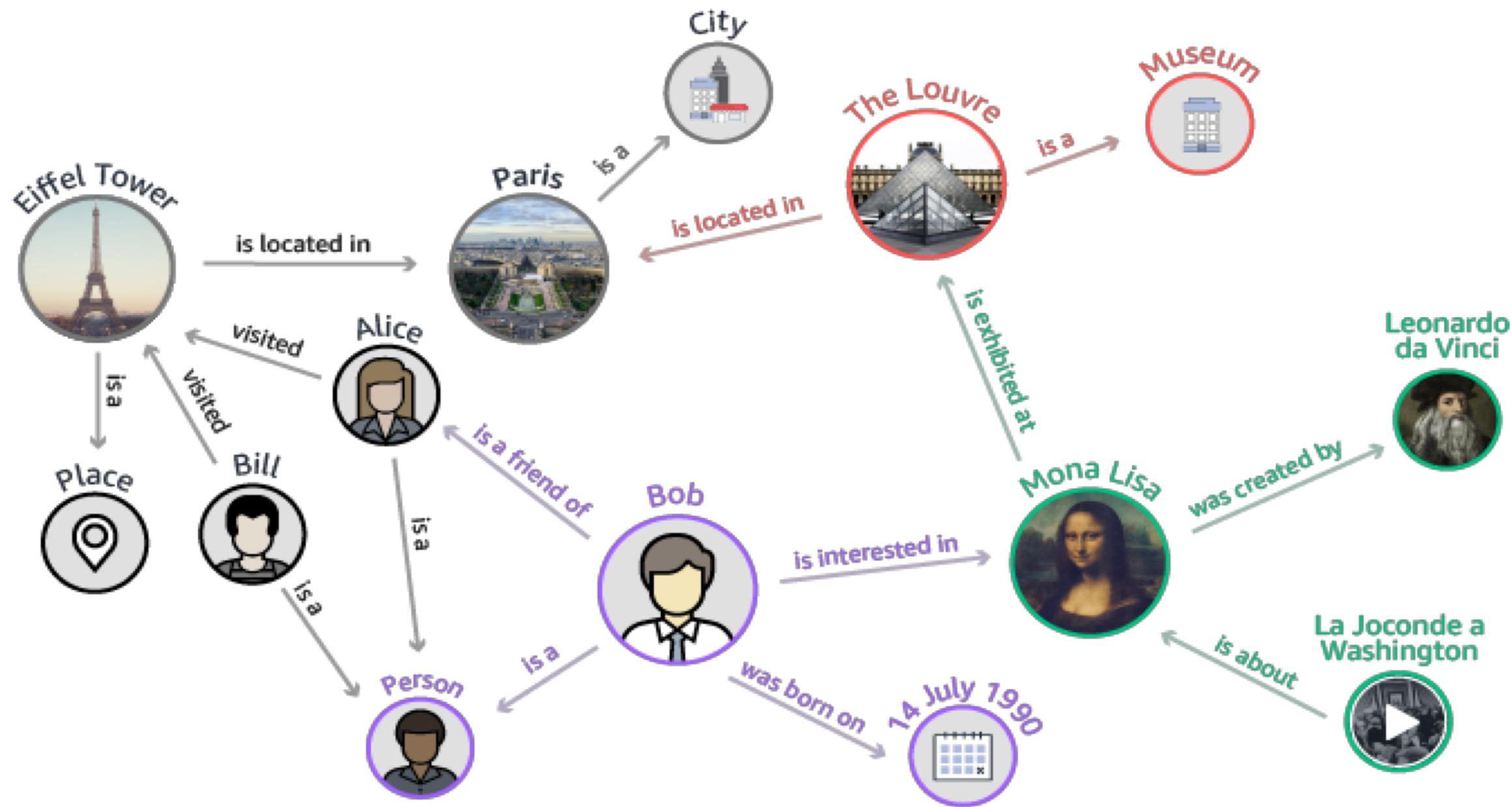
---

os nós (são os vértices do grafo)

os relacionamentos (são as arestas)

e as propriedades (ou atributos) dos nós e relacionamentos







# NoSQL – Grafos



# NoSQL – Grafos- Aplicação

---

Manipulação de dados estatísticos, frequentemente escritos mas raramente lidos

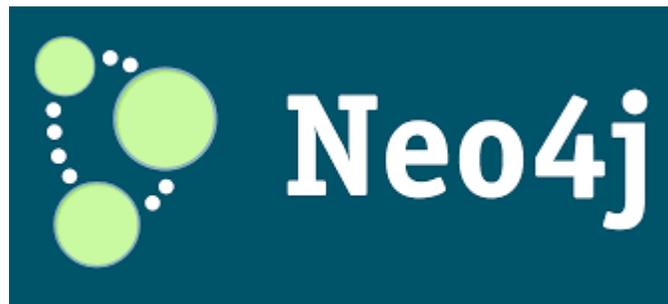
(como um contador de hits na Web)

Aplicações que exigem alto desempenho em consultas com muitas junções (Tabelas)

Redes Sociais, Recomendações (Foco em modelar a estrutura dos dados – interconectividade)

# NoSQL – Grafos - Uso

---



# SGBD's NoSQL



# 50 GB de dados para armazenar

---

## **Mysql (Relacional)**

300 ms – Write

350 ms – Read

## **Cassandra (NoSQL)**

0.12 ms – Write

15 ms – Read

## **Diferença:**

- Write – **2.500** vezes mais rápido o NoSQL
- Read – **23** vezes mais rápido o NoSQL

# Quando Usar?

<b>SGDB-Relacional</b>	<b>NoSQL</b>
	O armazenamento deve ser capaz de lidar com <b>carregamentos pesados</b>
Armazenamento é esperado para apresentar carregamento pesado também, mas consiste principalmente na <b>leitura</b> de operações	Você executa muitas operações de <b>escrita</b> no armazenamento
Você <b>prefere performance</b> a uma estrutura de dados sofisticada	Você quer um armazenamento que seja <b>escalável horizontalmente</b>
Você precisa de uma linguagem <b>SQL query poderosa</b>	Simplicidade é bom, como em uma <b>linguagem query bem simples</b> (sem joins)

# Panorama de uso - SGBD

Fonte: DB-Engines Ranking, solid IT

343 systems in ranking, February 2019

Rank			DBMS	Database Model	Score		
Feb 2019	Jan 2019	Feb 2018			Feb 2019	Jan 2019	Feb 2018
1.	1.	1.	Oracle	Relational, Multi-model	1264.02	-4.82	-39.26
2.	2.	2.	MySQL	Relational, Multi-model	1167.29	+13.02	-85.18
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	1040.05	-0.21	-81.98
4.	4.	4.	PostgreSQL	Relational, Multi-model	473.56	+7.45	+85.18
5.	5.	5.	MongoDB	Document	395.09	+7.91	+58.67
6.	6.	6.	IBM Db2	Relational, Multi-model	179.42	-0.43	-10.55
7.	7.	8.	Redis	Key-value, Multi-model	149.45	+0.43	+22.43
8.	8.	9.	Elasticsearch	Search engine, Multi-model	145.25	+1.81	+19.93
9.	9.	7.	Microsoft Access	Relational	144.02	+2.41	+13.95
10.	10.	11.	SQLite	Relational	126.17	-0.63	+8.89
11.	11.	10.	Cassandra	Wide column	123.37	+0.39	+0.59
12.	13.	17.	MariaDB	Relational, Multi-model	83.42	+4.60	+21.77
13.	12.	13.	Splunk	Search engine	82.81	+1.39	+15.55
14.	14.	12.	Teradata	Relational	75.97	-0.22	+2.98
15.	15.	18.	Hive	Relational	72.29	+2.38	+17.23



# Histórico (Ano 2.000 – Dias Atuais)

---

- Machine Learning – Deep Learning
- Big Data
- Data Science
- NoSQL (SGBD não Relacionais)
- Engenharia de Dados (Hadoop, Spark)

