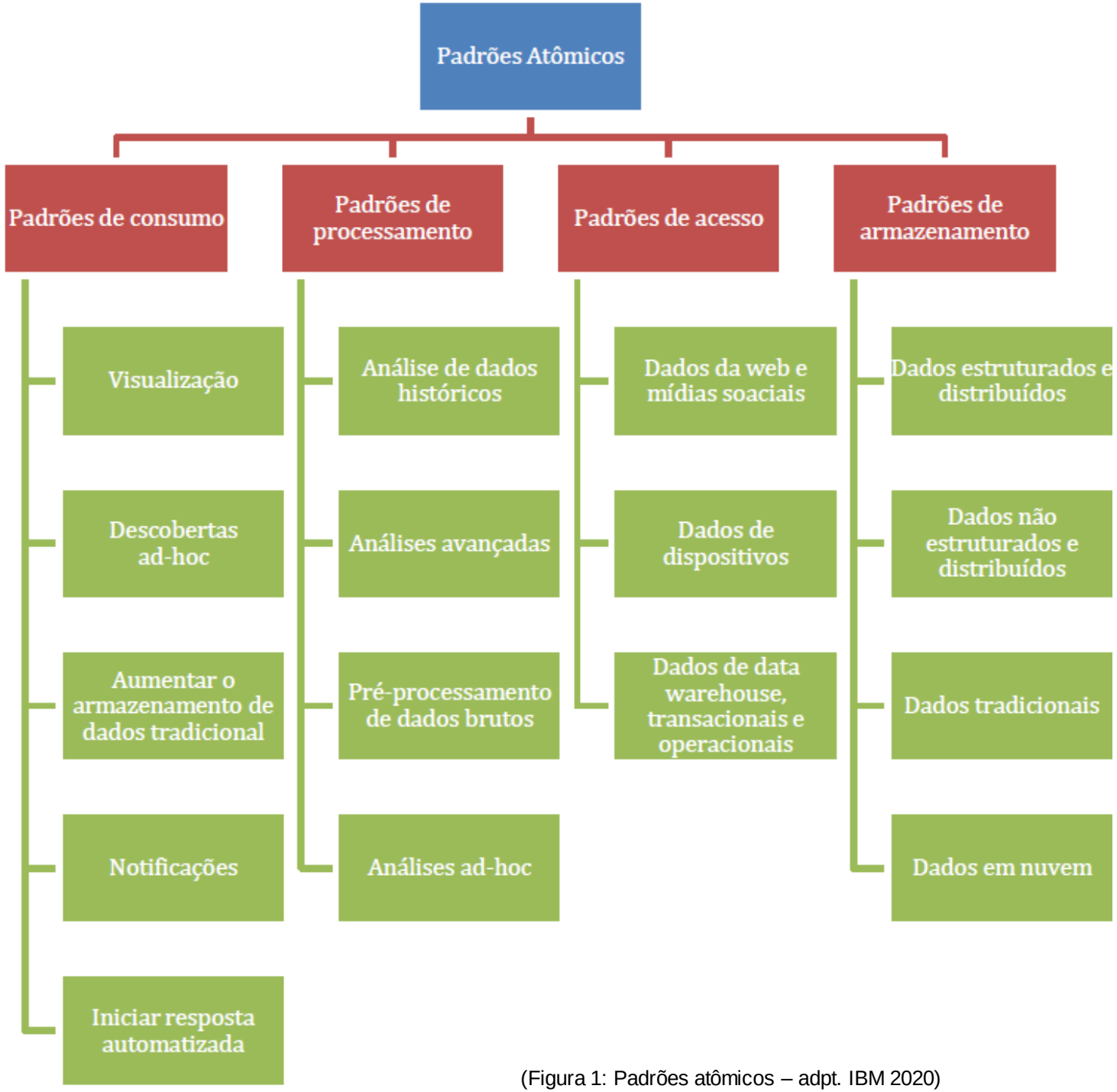
Capturando e Armazenando os Dados

Profa. Leticia T. M. Zoby

(leticia.zoby@udf.edu.br)



(Figura 1: Padrões atômicos – adpt. IBM 2020)

Padrões atômicos

- Ajudam a identificar a forma que os dados são consumidos, processados, armazenados e acessados por problemas de big data.
- Também ajudam a identificar os componentes necessários

Formas de Capturar Dados

- Objetivo: apresentar como cada tipo de dado requer uma estratégia diferente para ser utilizado no projeto
- Exemplo:

Uma loja de varejo desenvolveu um aplicativo com as seguintes funcionalidades (Marquesone, 2016):

- Permitir a pesquisa e compra das centenas de produtos da empresa;
- Permitir que o cliente avalie um produto e verifique os comentários de outros clientes;
- Permitir que o cliente compartilhe as informações de produtos e listas de compras nas redes sociais.

Crescimento rápido de usuários – aplicativo:

- Insatisfação
- Queda de serviço
- Lentidão de processo

Formas de Capturar Dados

- **Dados:**
  - **Dados internos**
    - Aqueles que são da "empresa" e que possuem controle
    - Há uma lista de conjunto de dados:
      - **Dados de sistema de gerenciamento da empresa**: sistemas de gerenciamento de projetos, automação, marketing, sistema CRM (*Customer Relationship Management*), sistema ERP (*Enterprise Resource Planning*), entre outros.
      - **Arquivos**: documentos escaneados, correspondências, notas fiscais, entre outros.
      - **Documentos gerados pelos colaboradores**: planilhas em formato XML, relatórios em formato PDF, dados em formatos CSV e JSON e-mails, entre outros.
      - **Sensores**: dados de medidores inteligentes, sensores de carros, câmeras de vigilância, entre outros.
      - **Registros de logs**: logs de eventos, dados de servidores, logs sobre uso de aplicativos móveis, entre outros.

Formas de Capturar Dados

- **Dados:**
  - **Dados sensores**
    - Solução inserida no contexto de IoT (*Internet of Things* - Internet das Coisas).
    - Objetos possuem capacidade de comunicação com outros objetos e pessoas.
    - Identificar um meio de transmissão de dados entre os sensores e um servidor para prover o armazenamento.
    - Exemplo: Uma empresa na área de logística e transporte e que sua frota possuem sensores e que geram informações. Essas podem ser gasto de combustível, qualidade da direção e estados dos pneus.

Formas de Capturar Dados

- Dados:
  - Dados sensores

Tecnologia	Característica	Aplicação
Bluetooth	Comunicação econômica usada para transmissão de dados em pouco alcance	Comunicação contínua entre os dispositivos e aplicações
Celular (2G, 3G, 4G)	Serviços de telefonia móvel para a comunicação entre uma ou mais estações	Atividades gerais na internet
NFC	Comunicação por campo de proximidade (Near Field Communication) que permite a troca de informações sem fio e de forma segura	Pagamentos e captura de informações de produtos
Wifi	Comunicação que permite a transmissão de dados em alta velocidade em diversas distâncias	Uso intensivo dos dados como streaming, Voip e download
Zigbee	Comunicação entre dispositivos com baixa potência de operação, baixa taxa de transmissão de dados e baixo custo	Aplicações que exigem baixo consumo de energia e baixas taxas de dados

Formas de Capturar Dados

- **Dados:**
  - **Dados da Web**
    - Dados coletados de fontes externas, com o propósito de verificar quais poderiam ser relevantes no projeto de Big Data. Há:
      - **Dados de domínio público**: dados disponibilizados pelo governo, dados sobre o clima,...
      - **Dados de sites de terceiros**: imagens, vídeos, áudios, podcasts, ...
      - **Mídias sociais online**: twitter, youtube, instagram,...
    - Como capturar esses dados?
      - Através da API (*Application Programming Interfce*): conjunto de instruções e padrões de programação
        - Ex:

          Facebook — https://developers.facebook.com/

          Flickr — https://www.flickr.com/services/api/

          Instagram — https://www.instagram.com/developer/

          LinkedIn — https://developer.linkedin.com/

          Pinterest — https://developers.pinterest.com/

          Twitter — https://dev.twitter.com/

          YouTube — https://developers.google.com/youtube/

Formas de Capturar Dados

- **Dados:**
  - **Dados abertos**
    - Dados de domínio público
      - Aceleração do desenvolvimento de pesquisas, como exemplo dados sobre a economia do país.
      - Aumento de qualidade da informação para a sociedade.

Formas de Capturar Dados

- Como todos esses dados disponíveis (dados internos, de sensores, da Web e abertos), é preciso uma estratégia para armazenar toda essa variedade!! E quais os requisitos para o armazenamento desses dados????
- Desafios enfrentados:
  - Escalabilidade
  - Alta Disponibilidade
  - Flexibilidade

Formas de Capturar Dados

- Desafios enfrentados:
  - **Escalabilidade**
    - quantidade de dados pode crescer aceleradamente à medida que novos usuários e funcionalidades são adicionados à solução. Essa solução é considerada escalável se ela for capaz de manter o desempenho desejável mesmo com a adição de nova carga.
    - Os SGBDRs -> escalabilidade vertical

(SGBDR = Sistema de Gerenciamento de Banco de Dados Relacionais)

Formas de Capturar Dados

- Desafios enfrentados:
  - **Alta Disponibilidade**
    - Precisa manter o serviço disponível.

    Ex: carrinho de compra, mesmo havendo uma inconsistência nas informações do pedido do cliente, de forma que ele não liste todos os produtos que o cliente selecionou, é melhor garantir que o serviço continue disponível e o cliente precise atualizar seu pedido do que interromper o serviço, impedindo-o de finalizar sua compra.

Formas de Capturar Dados

- Desafios enfrentados:
  - **Flexibilidade**
    - Os SGBDRs necessita de toda modelagem dos dados antes de armazená-los. E em soluções atuai isso não é possível, não há conhecimento antecipado.

- Para suprir esses requisitos, novas alternativas foram desenvolvidas, nascendo o termo NoSQL.

Tecnologia NoSQL

- NoSQL é uma abreviação de *Not only SQL*, ou seja "não somente SQL".
- Esse termo foi cunhado para definir os novos modelos de armazenamento de dados, criados para atenderem às necessidades de flexibilidade, disponibilidade, escalabilidade e desempenho das aplicações inseridas no contexto de Big Data.

Tecnologia NoSQL

- O objetivo do NoSQL não é substituir a linguagem SQL. É usar também modelos não-relacionais, para trazer a melhor solução para um determinado problema.
- Características comuns dos BD NoSQL:
  - Não utilizam o modelo relacional;
  - Tem uma boa execução em cluster;
  - Ter código aberto;
  - Criados para suportar propriedades da web do século XXI; e
  - Não tem um esquema definidos.

- De acordo com a estrutura que os dados são armazenados, há 4 modelos principais:
  - **orientado a chave-valor,**
  - **orientado a documentos,**
  - **orientado a colunas; e**
  - **orientado a grafos.**

Tecnologia NoSQL

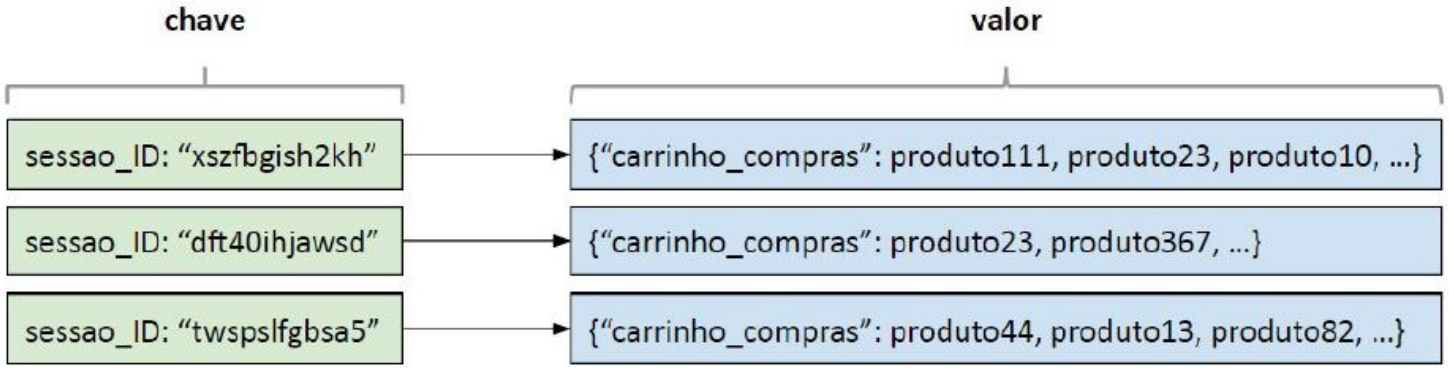
Modelo orientado a chave-valor

- É o mais simples.
- Como o nome já diz, esse tipo de armazenamento tem como estratégia o armazenamento de dados utilizando chaves como identificadores das informações gravadas em um campo identificado como valor.
- Baseado em tabelas de *hash* para garantir que cada registro seja armazenado com uma chave única.
- São adequados para aplicações que realizam leituras frequentes.

Tecnologia NoSQL

Modelo orientado a chave-valor

- Exemplo: Os clientes acessam o catálogo de produtos do aplicativo e selecionam os itens desejados para colocá-los no carrinho de compras. Nesse momento, a aplicação precisa guardar as informações dos produtos selecionados até o momento em que o cliente finalize sua compra.



(Figura 2: Estrutura de um BD orientado a chave-valor Marquesone, 2016)

Tecnologia NoSQL

Modelo orientado a chave-valor

- Exemplos de BD orientado a chave-valor:
  - DynamoDB — https://aws.amazon.com/pt/dynamodb/
  - Redis — http://redis.io/
  - Riak — http://basho.com/
  - Memcached — https://memcached.org/

Tecnologia NoSQL

Modelo orientado a documentos

- Como o próprio nome diz, este modelo armazena coleções de documentos.
- Um documento, em geral, é um objeto com um identificador único e um conjunto de campos, que podem ser *strings*, listas ou documentos aninhados. Estes campos se assemelham a estrutura chave-valor, porem tem um conjunto de documentos e em cada documento tem um conjunto de campos (chaves) e o valor deste campo.
- A diferença é que cada documento (ou seja, cada linha da tabela) pode conter variações em sua estrutura. Isso é possível pelo fato de que não é preciso definir um esquema antes de adicionar os registros.

Tecnologia NoSQL

Modelo orientado a documentos

- Exemplo:

```
{
    "clientes" :[
      {
        "primeiroNome" : "João",
        "ultimoNome" : "Silva",
        "idade" : 30,
        "email" :"xx@y.com",
        "fone" : "11-984592015"
      },
      {
        "primeiroNome" : "José",
        "ultimoNome" : "Pereira",
        "idade" : 28,
        "email" :"aaa@b.com",
        "contato" {
            "foneFixo" : "11-52356598",
            "foneCelular" : "11-987452154",
            "foneComercial" : "11-30256985"
        }
      }
    ]
}
```

(Figura 2: Estrutura de um BD orientado a documentos (Marques, 2016)

Tecnologia NoSQL

Modelo orientado a documentos

- Caso seja necessário uma solução que armazene atributos variados em cada registro, o banco de dados orientado a documentos é uma ótima opção. Além disso, ele oferece grande escalabilidade e velocidade de leitura.
- Permite trabalhar com a replicação dos dados em um cluster.
- Esse modelo é indicado para realizar o armazenamento de conteúdo de páginas Web, na catalogação de documentos de uma empresa e no gerenciamento de inventário de um e-commerce.

Tecnologia NoSQL

Modelo orientado a documentos

- Exemplos de BD orientado a documentos:
  - Couchbase — http://www.couchbase.com/
  - CouchDB — http://couchdb.apache.org/
  - MarkLogic — http://www.marklogic.com/
  - MongoDB — https://www.mongodb.com/

Tecnologia NoSQL

Modelo orientado a colunas

- É o mais complexo

- Voltado para computação distribuída, eficiente ao armazenar grandes quantidades de dados separando-os em muitas máquinas. Para isso, utiliza as famosas Big Tables, tabelas gigantes onde as chaves apontam para várias colunas distintas. Preferível para motores de buscas, como o Google.

Tecnologia NoSQL

Modelo orientado a colunas

- Exemplo:

CLIENTE	
ID_CLIENTE	INT(10)
NOME	VARCHAR(100)
IDADE	INT(3)
EMAIL	VARCHAR(100)
FONE	VARCHAR(10)

(Figura 4: Tabela Cliente de um BDRelacional (Marquesone, 2016)

- Agora, definir 3 famílias de colunas: dados_cadastrais, preferencia_roupas e preferencia_livros.
  - A partir delas, o desenvolvedor possui a flexibilidade de inserir as colunas que considerar necessárias em cada registro armazenado, sem precisar alterar a estrutura dos dados já armazenados.

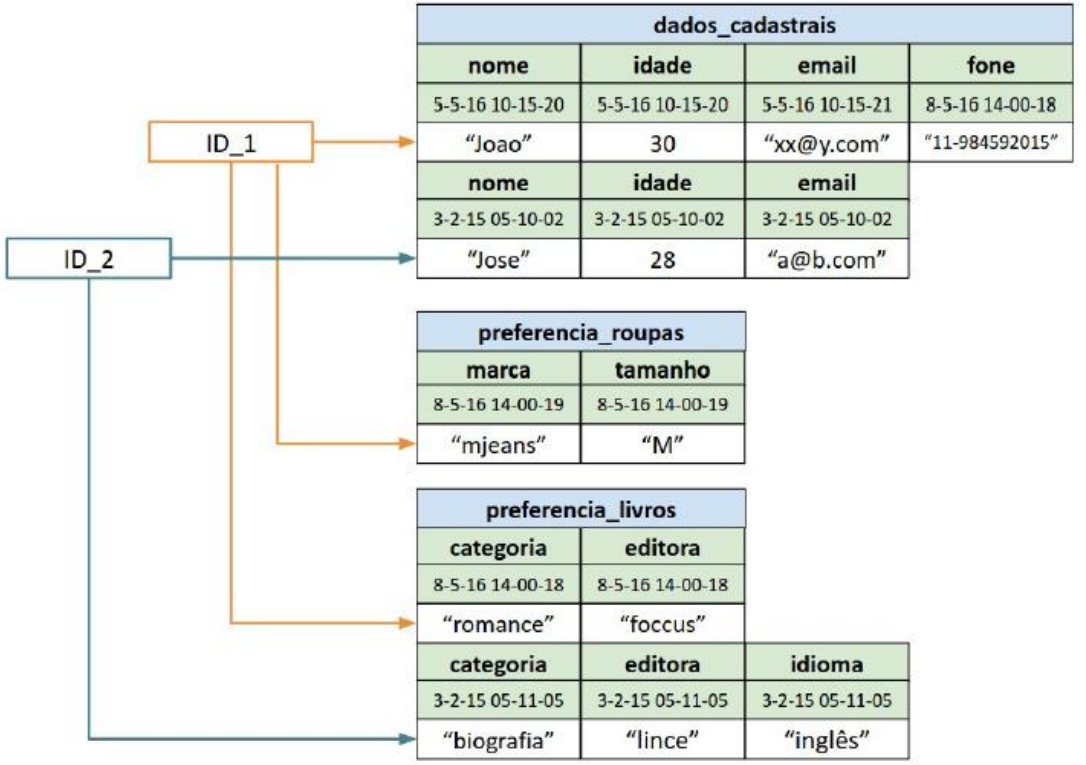
Tecnologia NoSQL

Modelo orientado a colunas

- Exemplo:

Os dados são armazenados fisicamente em uma sequência orientada a colunas e não por linhas.

- A sequência seria: "João", "José", 30, 28, xx@y.com, a@b.com, ...



(Figura 5: Exemplo de família de colunas (Marquesone, 2016)

Tecnologia NoSQL

Modelo orientado a colunas

- Exemplos de BD orientado a colunas:
  - Accumulo — https://accumulo.apache.org/
  - Cassandra — http://cassandra.apache.org/
  - HBase — https://hbase.apache.org/
  - Hypertable — http://www.hypertable.org/

Tecnologia NoSQL

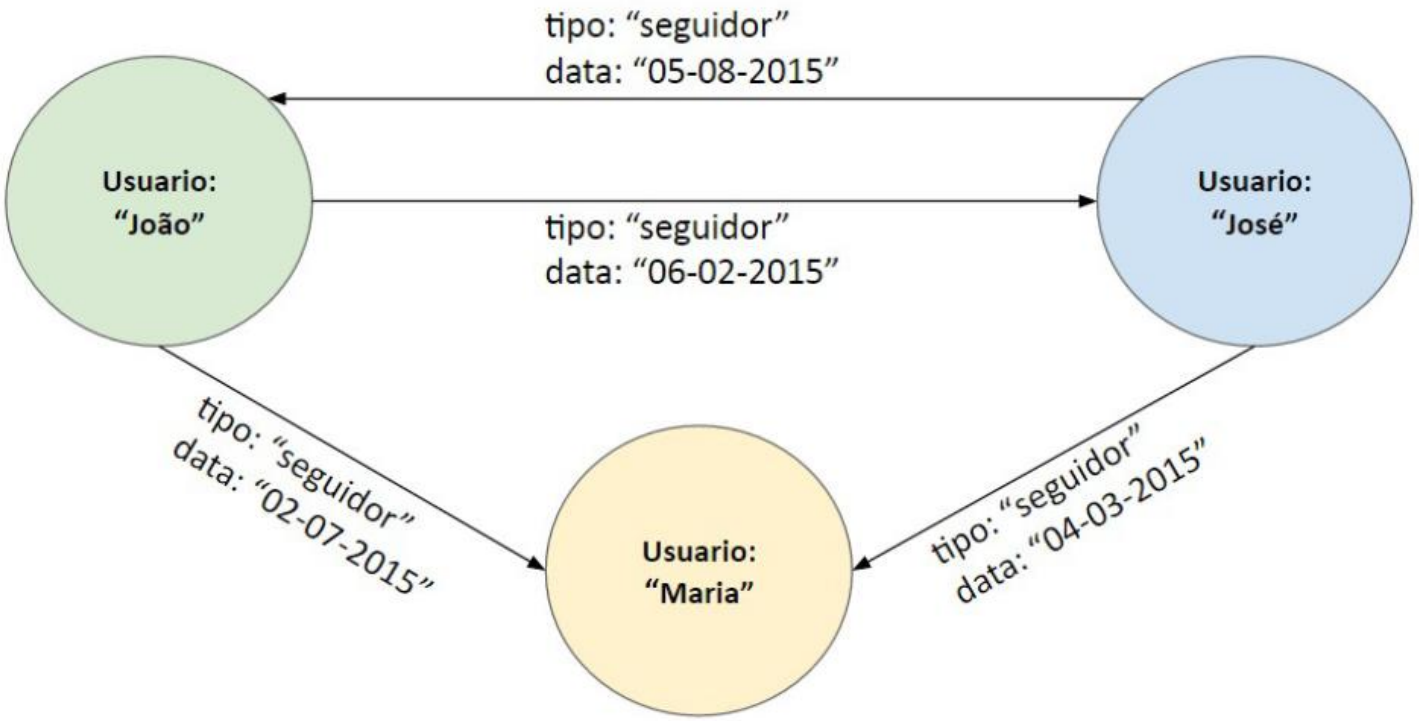
Modelo orientado a grafos

- É o mais especializado, pois possuem uma estrutura definida na teoria dos grafos, usando vértices e arestas para armazenar os dados dos itens coletados (como pessoas, cidades, produtos e dispositivos).
- Baseado em relacionamentos entre nós, os quais possuem possibilidade de mudança de formato individual.
- Torna-se fácil fazer conexões entre eles, assemelhando-se ao conceito de banco com registros encadeados.

Tecnologia NoSQL

Modelo orientado a grafos

- Sua estrutura é ideal para a modelagem de redes sociais.



(Figura 6: Exemplo de BD orientados a grafos (Marquesone, 2016)

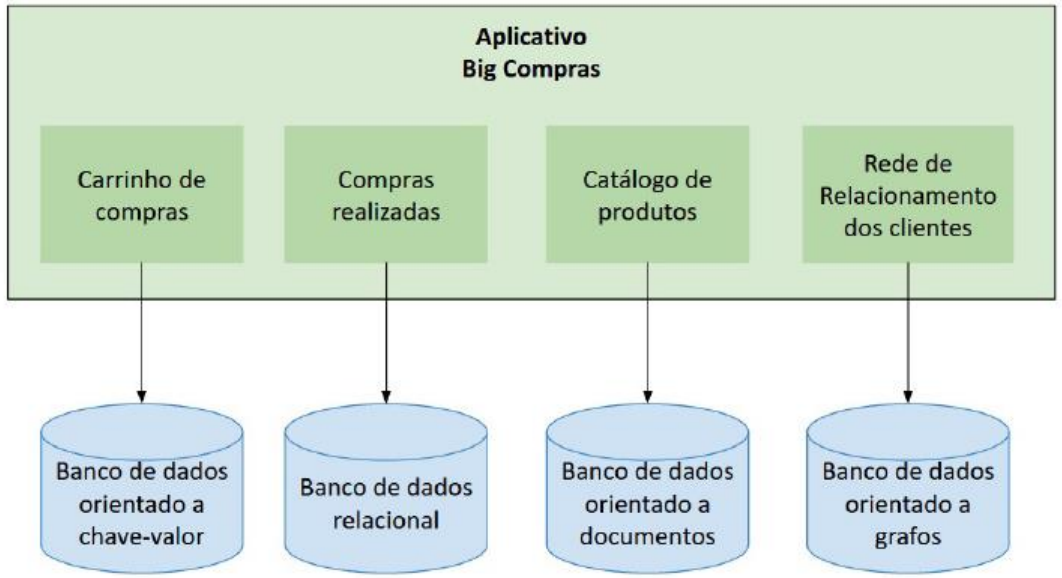
Tecnologia NoSQL

Modelo orientado a grafos

- Exemplo:
- Exemplos de BD orientado a grafos:
  - AllegroGraph — http://franz.com/agraph/allegrograph/
  - ArangoDB — https://www.arangodb.com/
  - InfoGrid — http://infogrid.org/trac/
  - Neo4J — https://neo4j.com/
  - Titan — http://titan.thinkaurelius.com/

Tecnologia NoSQL

Exemplo: aplicativo de compras de uma empresa na área de varejo (Marquesone, 2016):

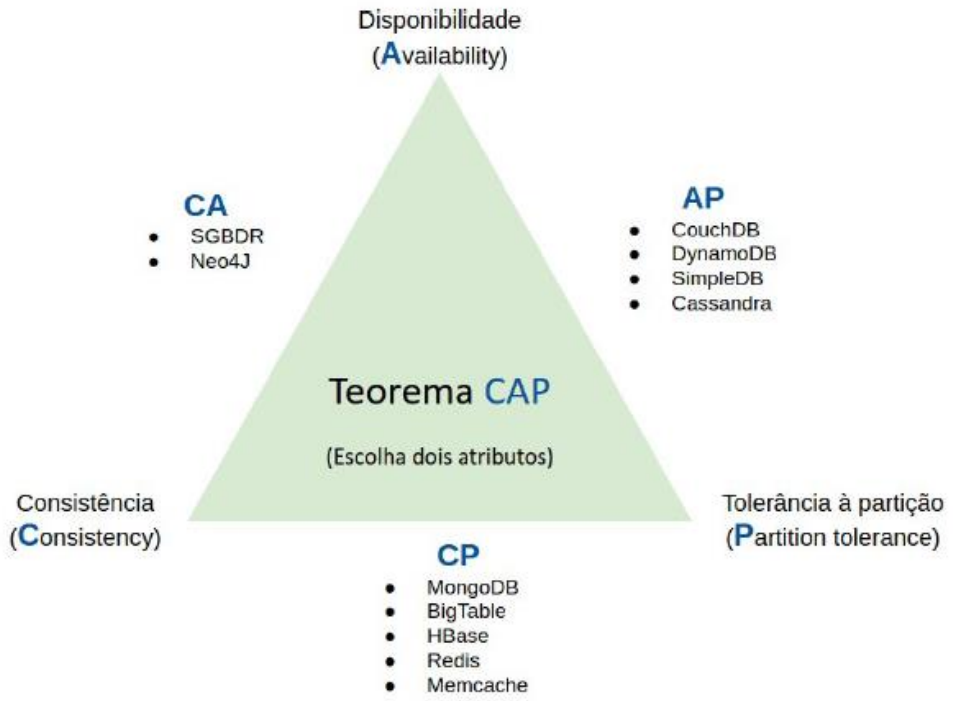


(Figura 7: Exemplo de solução híbrida de armazenamento de dados (Marquesone, 2016)

- Desafio: Decidir qual o melhor banco de dados para cada serviço.

Tecnologia NoSQL – Teorema CAP

- Em 2000 foi proposto por Eric Brewer o teorema CAP para BD NoSQL.
- O teorema consiste no seguinte conjunto de requisitos para sistemas distribuídos: consistência (*Consistency*), disponibilidade (*Availability*) e tolerância à partição (*Partition tolerance*).



(Figura 8: Teorema CAP (Marquesone, 2016)

Tecnologia NoSQL - Teorema CAP

- Segundo Brewer, é teoricamente impossível obter um sistema que atenda os 3 requisitos.
  - **Consistência**: refere-se ao aspecto que todos os nós do sistema devem conter os mesmos dados, garantindo que diferentes usuários terão a mesma visão do estado dos dados.
  - **Disponibilidade**: o sistema deverá sempre responder a uma requisição, mesmo que não esteja consistente.
  - **Tolerância à partição**: deve garantir que o sistema continuará em operação mesmo que algum servidor do cluster venha a falhar.

Referência

IBM (2020). **Entendendo padrões atômicos e compostos de soluções de big data.** Disponivel em: https://www.ibm.com/developerworks/br/library/bd-archpatterns4/ Acessado em: 20 abr. 2021.

MARQUESONE, Rosangela. **Big Data. Técnicas e tecnologia para extração de valor dos dados**. Casa do Código. 2017.

Referências...