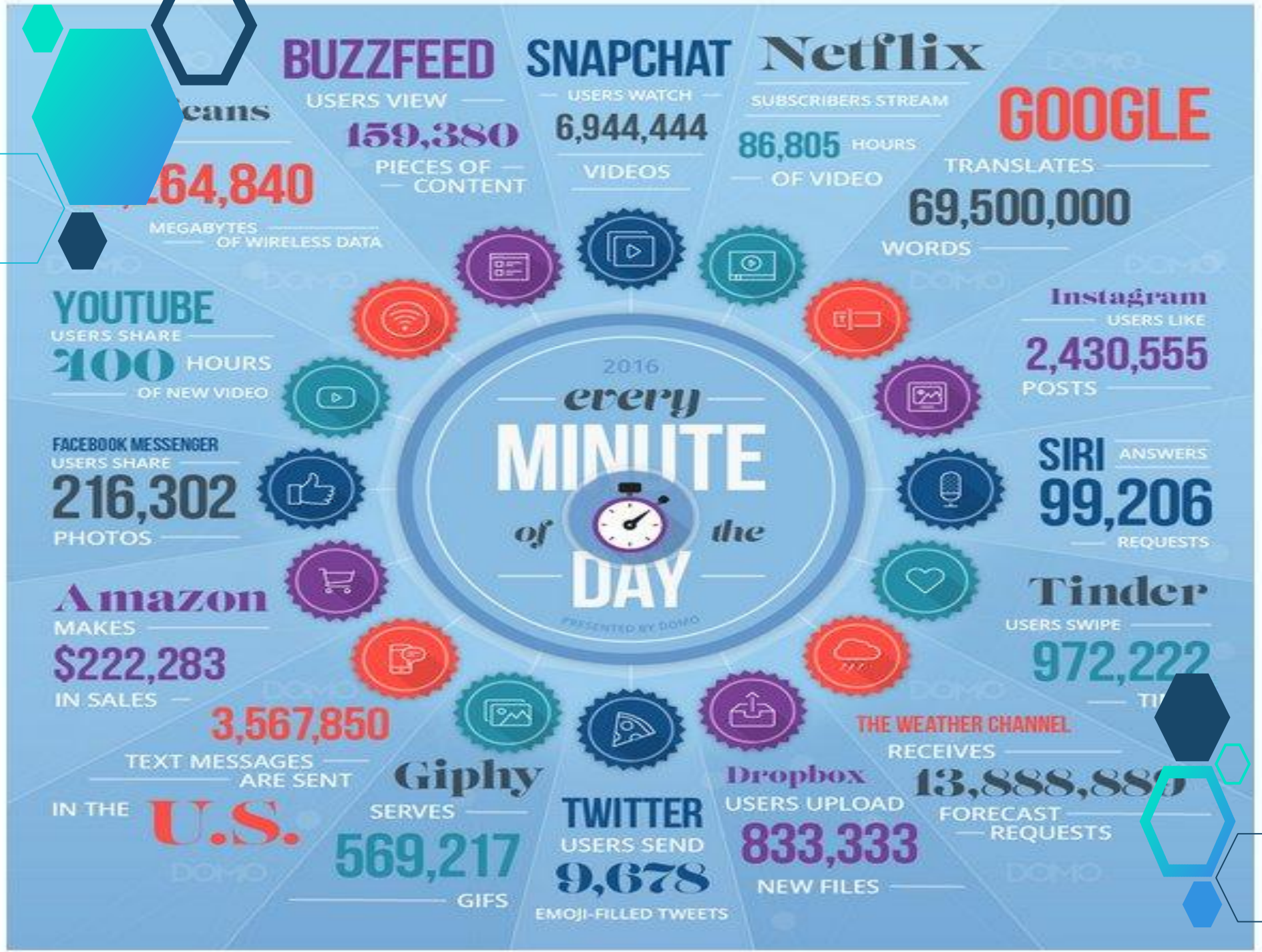
Big Data Integration
Alinhamento de Esquemas

Helton Santos

Roteiro

- Big Data
  - 5 V's
  - Ciclo de Vida
  - Soluções para Big Data
- Integração de Dados
  - Alinhamento de Esquemas
- Big Data Integration
  - Enfrentando o desafio da variedade e velocidade
  - Enfrentando o desafio da variedade e volume
    - Deep Web
    - Web Tables

1 Big Data


BUZZFEED
USERS VIEW
159,380
PIECES OF CONTENT
SNAPCHAT
USERS WATCH
6,944,444
VIDEOS
Netflix
SUBSCRIBERS STREAM
86,805 HOURS
OF VIDEO
GOOGLE
TRANSLATES
69,500,000
WORDS
MEGABYTES OF WIRELESS DATA
YOUTUBE
USERS SHARE
400 HOURS
OF NEW VIDEO
Instagram
USERS LIKE
2,430,555
POSTS
2016
every
MINUTE
of the
DAY
PRESENTED BY DOMO
FACEBOOK MESSENGER
USERS SHARE
216,302
PHOTOS
SIRI ANSWERS
99,206
REQUESTS
Amazon
MAKES
$222,283
IN SALES
Tinder
USERS SWIPE
972,222
3,567,850
TEXT MESSAGES
ARE SENT
IN THE U.S.
THE WEATHER CHANNEL
RECEIVES
13,888,888
FORECAST
REQUESTS
Giphy
SERVES
569,217
GIFS
TWITTER
USERS SEND
9,678
EMOJI-FILLED TWEETS
Dropbox
USERS UPLOAD
833,333
NEW FILES

“Big Data, em geral, é definido como ativos de alto volume, velocidade e variedade de informação que exigem custo-benefício, de formas inovadoras de processamento de informações para maior visibilidade e tomada de decisão.” GartnerGroup (2012)

“Big Data é o termo utilizado para descrever grandes volumes de dados e que ganha cada vez mais relevância à medida que a sociedade se depara com um aumento sem precedentes no número de informações geradas a cada dia”. IBM (2014)


BIG
DATA

“Big Data é um novo termo usado para identificar os conjuntos de dados que devido ao seu grande tamanho e complexidade, nós não podemos gerenciá-los com as metodologias e tecnologias tradicionais”. [Wei Fan e Albert Bifet, 2012].

5 V's

- **Volume**: Refere-se ao número de dados que serão manipulados e posteriormente analisados.
- **Velocidade**: Refere-se à velocidade em que os dados são gerados, armazenados e processados.
- **Variedade**: Refere-se aos diversos tipos de dados provenientes de varias fontes, sendo eles estruturados ou não estruturados.
- **Veracidade**: Refere-se à confiabilidade e precisão dos dados obtidos.
- **Valor**: Refere-se à vantagem em que o dado pode oferecer em sua análise.

Ciclo de Vida

Fase de geração

Os dados são criados em diferentes fontes

Fase de aquisição

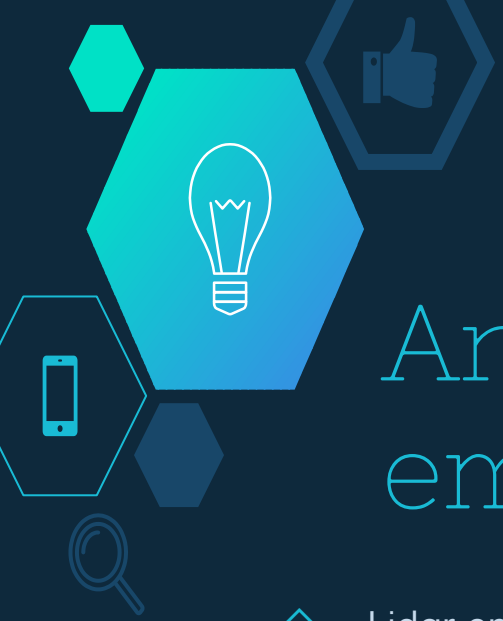
Os dados são coletados, transmitidos e pré-processados.

Fase de armazenamento

Os dados são armazenados em repositórios escaláveis e tem condições de gerenciar o grande volume de dados

Fase de análise

Aqui são extraídos insights dos dados, bem como tendências de mercado, comportamento de consumidores e suas expectativas.



Armazenamento em Big Data

- Lidar em a escalabilidade e elasticidade
- SGBDs relacionais tradicionais não suportam aplicações com volumes de dados que crescem exponencialmente em pouco tempo
- Chegada do **NoSQL**, que possibilita o armazenamento dos dados de diversas formas
- Para enormes volumes de dados, escalar dentro do NoSQL é uma uma tarefa fácil e menos custosa.
- São otimizados para trabalhar com processamento paralelo, distribuição global e aumento imediato da sua capacidade.

Armazenamento em Big Data

2 Integração de Dados

“A motivação para criação de um sistema de integração de dados é oferecer aos usuários uma interface uniforme de acesso à diferentes fontes de dados autônomas.”

Abordagens da integração de dados

Virtual

Os dados são buscados nas fontes apenas quando as consultas são requisitadas.

Os dados estão sempre atualizados porém torna-se ineficiente se suas fontes ficarem inacessíveis, ou se o processamento de tradução e integração for muito demorado.

Materializada

Os dados são obtidos, tratados e armazenados num repositório central.

Os dados estão disponíveis imediatamente para consultas, porém é preciso manter a consistência entre os dados armazenados e os das fontes de origem.

Desafios na Integração de Dados

Modelagem de Dados

- Heterogeneidade das fontes
- Possuem diferentes denominações para os mesmos conceitos
- Quando se envolve muitas fontes se torna inviável

Reformulação de Consultas

- É importante garantir que fontes de dados irrelevantes não estejam sendo acessadas
- Velocidade de execução das consultas

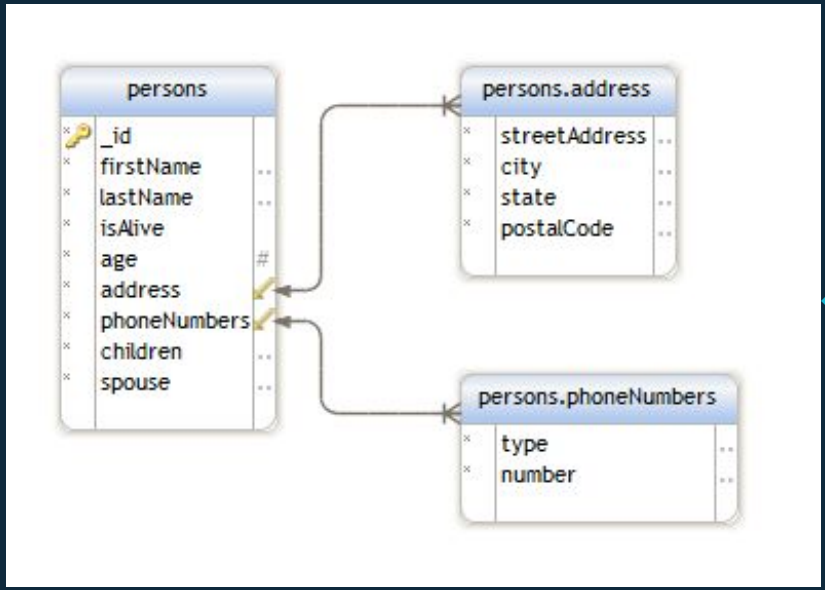
Manutenção dos Sistemas de Integração

- Fontes autônomas
- Sofrem alterações constantes, tanto nos dados quanto nos esquemas.

2.1 Alinhamento de Esquemas

O que é um esquema de banco de dados?

Esquema de banco de dados



- É uma estrutura descrita em uma linguagem formal e refere-se à organização e construção dos dados em um banco de dados.
- Um esquema de banco de dados é especificado por um conjunto de definições expressas por uma linguagem especial chamada linguagem de definição de dados (Data Definition Language, DDL).

Alinhamento De esquemas

- Aborda desafio da ambiguidade semântica
- Visa compreender quais atributos têm o mesmo significado entre os esquemas

Esquema mediador

- Prover uma visão unificada e virtual dos esquemas heterogêneos e autônomos
- Captura detalhadamente os aspectos de domínio que está sendo considerado
- Muitas vezes é criado manualmente

Matching de atributos

- Prover um match dos atributos de cada esquema heterogêneo
- Esse match corresponde com os atributos do esquema mediador
- Algumas vezes, um atributo do esquema mediador pode corresponder à combinação de vários atributos dos esquemas heterogêneos

Mapeamento de esquemas

- Prover um mapeamento entre cada esquema heterogêneo com o esquema mediador, especificando os relacionamentos semânticos de cada atributo
- O resultado do mapeamento de esquemas é utilizado para reformular consultas submetidas pelos usuários nas fontes de dados subordinadas, tomando como base o esquema mediador

Desafios

- Não é uma tarefa trivial
- Diferentes fontes podem descrever o mesmo domínio usando esquemas muito diferentes
- Eles podem usar nomes de atributos distintos mesmo quando possuem o mesmo significado
- Fontes de dados podem ter atributos com o mesmo nome, mas com significados diferentes

3 Big Data Integration

“Inicialmente, é importante resaltar que, se tratando de big data, o alinhamento de esquemas se torna uma tarefa ainda mais complicada. Em vez de integrar dados de uma organização, o objetivo muitas vezes é integrar dados estruturados ou semi estruturados a partir da web, seja na deep web, ou em formulários, tabelas e listas.”

Desafios

- Enorme volume dados
- Grande variedade de dados
- Mudança constante de seus esquemas
- Uma fonte que está disponível hoje, amanhã poderá não mais existir
- Manter o mapeamento dos esquemas atualizado é uma tarefa que se torna inviável
- Atrasos na construção da integração

3.1 Desafio da variedade e velocidade

Possível solução

- Franklin et al. propuseram uma plataforma que da suporte ao mapeamento de esquemas em 2005
- Essa plataforma aborda o desafio da variedade e velocidade no contexto de big data
- Utilizando o paradigma pay-as-you-go, ela constrói os mapeamentos de esquemas entre diferentes fontes de dados conforme a necessidade

A plataforma

- Dada uma consulta, essa plataforma gera respostas aproximadas das fontes de dados, mesmo onde o mapeamento de esquemas ainda não é consolidado
- A medida que são feitas consultas e atividades de mineração nessas fontes de dados, a plataforma orienta e auxilia os usuários a construir a integração dessas fontes de maneira mais precisa, melhorando assim o alinhamento dos esquemas

3.2

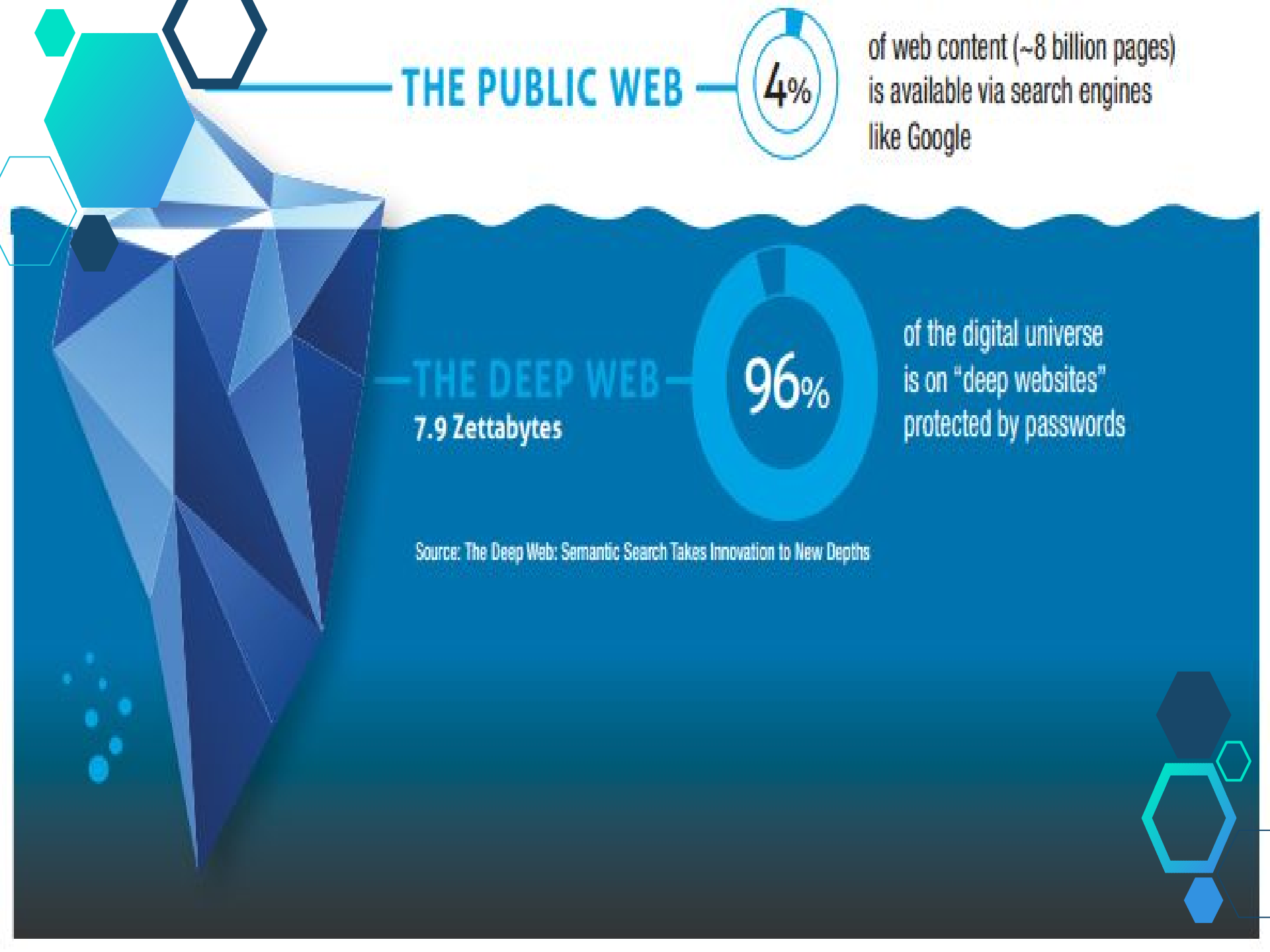
Desafio da variedade e volume

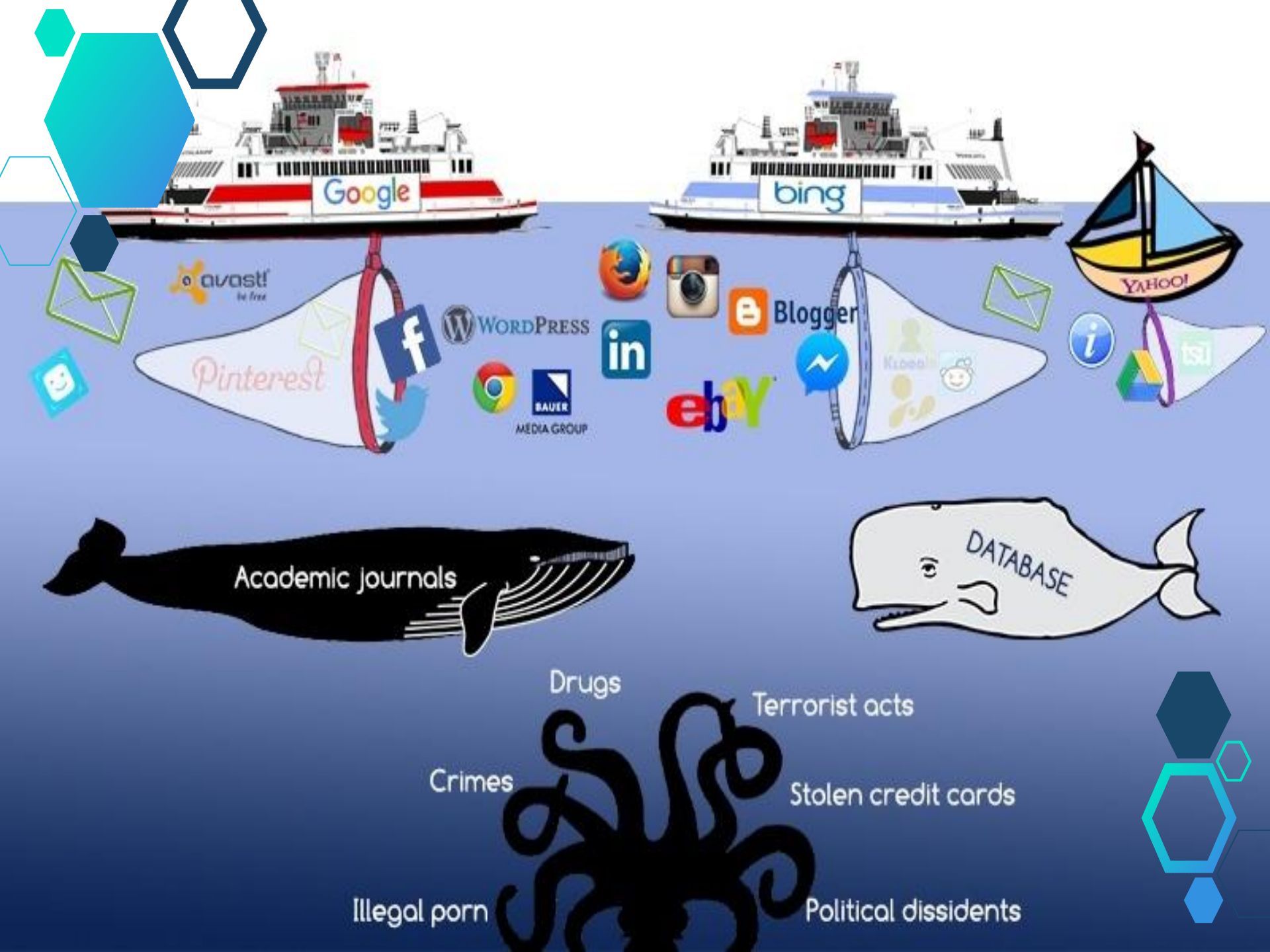
Neste ambiente de grande volume e variedade de dados, podemos destacar dois tipos de estruturas de dados na web, a Deep Web e as Web Tables.

3.2.2 Deep Web

Deep Web

- É a parte da Web que é formada por inúmeros sites e conteúdos que não são acessíveis por link padrões.
- Seu conteúdo não pode ser indexado por buscadores comuns
- Estima-se que a deep web seja 500 vezes maior do que a web comum


THE PUBLIC WEB
4%
of web content (~8 billion pages)
is available via search engines
like Google
THE DEEP WEB
7.9 Zettabytes
96%
of the digital universe
is on "deep websites"
protected by passwords
Source: The Deep Web: Semantic Search Takes Innovation to New Depths


Google
bing
avast!
Pinterest
WordPress
Blogger
YAHOO!
BAUER
MEDIA GROUP
ebay
Academic journals
DATABASE
Drugs
Terrorist acts
Crimes
Stolen credit cards
Illegal porn
Political dissidents

Possível solução

- Oferecer acesso unificado aos dados
- Construir um matching de esquemas holístico sobre formulários na web
- Com o matching é possível descobrir um esquema central, a partir disso construir um esquema de domínio completo que melhor descreve todas as instâncias dos dados
- Cada atributo do esquema, está em um grupo de atributos com mesmo significado semântico

1 de 10

Notebook Asus X555UB Intel Core i7 6500U 15,6" 8GB HD 1 TB GeForce 940M

★★★★★ 1 avaliação

a partir de
R$ 2.699,99
Ou 10x R$ 299,99
em 3 lojas

Salvar | Alerta de Preço | Comparar com outro

Intel Core i7 | 6ª Geração | Windows 10 Home | Wi-Fi | Série X | Velocidade do Processador 2,5 GHz

"Esse notebook Asus é ideal para quem precisa trabalhar com programas mais pesados, ou quer abrir games de última geração, principalmente por conta de seu..."

Ver análise completa

Encontramos preços em 3 lojas confiáveis

Frete para: 44927-000 ALTERAR

Ordenar

	Submarino	R$ 2.699,99 à vista ou 10x de R$ 299,99	Frete: valor não informado	Ir à loja
	americanas.com	R$ 2.893,49 à vista ou 10x de R$ 321,49	Frete: valor não informado	Ir à l
	shoptime	R$ 2.893,49 à vista ou 10x de R$ 321,49	Frete: valor não informado	Ir à loja

MELHOR PREÇO GARANTIDO

Voo + Hotel com o melhor preço garantido

Pacotes

Economize até 35%

Voo + Hotel | Voo + Carro | Hotel + Carro

Origem
Recife, Brasil

Destino
Brasília, Brasil

Quando?
01 Dez 2016 — 15 Dez 2016

Ainda não defini as datas

Quartos
1

Adultos
1

Crianças
0

Opções avançadas

Procurar

Refinar a sua pesquisa

ESTRELAS

Ordenar por Mais Vendidos

8.8 Meliá Brasil 21

★★★★★ 4,0 km do centro | Ver mapa

Últimos 2 quartos

IDA Qui 1 Dez 2016	Sai: 05:33	REC	BSB	Chega: 09:10	Direto
VOLTA Qui 15 Dez 2016	Sai: 19:40	BSB	REC	Chega: 21:12	Direto

Tam

Modificar voo

Economize R$ 116

Hotel + voo
~~R$ 5.521~~
R$ 5.405
Por pessoa

1 pessoa R$ 5.405
TOTAL: R$ 5.405

A incluir taxas e encargos

Ver hotel

Compre em parcelas!

8.5 Nobile Suítes Monumental

★★★★ 4,3 km do centro | Ver mapa

Últimos 3 quartos

IDA Qui 1 Dez 2016	Sai: 05:33	REC	BSB	Chega: 09:10	Direto
VOLTA Qui 15 Dez 2016	Sai: 19:40	BSB	REC	Chega: 21:12	Direto

Economize R$ 116

Hotel + v...
~~R$ 5.7...~~
R$ 5.589
Por pessoa

1 pessoa R$ 5.589
TOTAL: 5.589

A incluir taxas e enca...

Ver hotel

3.2.2 Web Tables

Web Tables

- São dados relacionais em formato de tabelas na Web
- Essa estrutura de dados é diferente da deep web, elas são tabelas HTML que já são rastreáveis, sem precisar preencher qualquer formulário.
- As Web Tables não tem esquemas claramente especificados, e sua semântica é frequentemente descrita na coluna da tabela e as vezes podem não ter

Fi[illegible]	Released	Studio	Worldwide Gross	Domestic Gross
Fr[illegible]	2013	Disney	$1,072,402,000	$398,402,000
Toy Story 3	2010	Disney Pixar	$1,063,171,911	$415,004,880
The Lion King	1994	Disney	$987,483,777	$422,783,777
Despicable Me 2	2013	Universal	$970,761,885	$368,061,265
Finding Nemo	2003	Pixar	$936,743,261	$380,843,261
Shrek 2	2004	Dreamworks	$919,838,758	$441,226,247
Ice Age: Dawn of the Dinosaurs	2009	Fox	$886,686,817	$196,573,705
Ice Age: Continental Drift	2012	Fox	$877,244,782	$161,[illegible]43
Shrek the Third	2007	Dreamworks	$798,958,162	$322,[illegible]719,[illegible]44
Shrek Forever After	2010	Dreamworks	$752,600,867	$238,73[illegible]787

Soluções de recuperação e integração de dados nas Web Tables

Primeiro

É a busca por palavras-chaves em web tables, onde o objetivo é aceitar consultas de palavras-chaves submetidas pelo usuários e classificar as tabelas por relevância.

Segundo

É encontrar tabelas relevantes, onde o objetivo é retornar tabelas que têm dados semelhantes ou complementares à tabela editada pelo usuários, para possivelmente fornecer referência.

Terceiro

É extrair conhecimento das web tables, onde o objetivo é extrair (entidade, propriedade e valor) triplas que podem ser usadas para povoar bases de conhecimento.

Considerações Finais

- Embora existam abordagens e plataformas que tratam da integração de dados em Big Data, isto ainda não é algo tão trivial de se fazer
- Ainda existe um amplo espaço para o desenvolvimento e pesquisa de novas tecnologias que facilitem a integração destas fontes de dados de forma mais automatizada

Referências

- X. L. Dong and D. Srivastava. Big Data integration. Synthesis Lectures on Data Managment. March 2015
- AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. Principles of Data Integration. MorganKaufmann, 2012.2
- M. Chen, S. Mao, and Y. Liu. Big Data: A Survey. Springer Science and Business Media New York, 2014
- X. L. Dong and D. Srivastava. Big Data Integration. 2013 IEEE 29th International Conference on Data Engineering (ICDE) (2013)
- Fan W., Bifet A. Mining big data: current status, and forecast to the future. ACM SIGKDD Explor Newsl, 2012.
- I. Taleb, D. Rachida and S. Mohamed Adel. Big Data Pre-Processing: A Quality Framework, 2015.

Referências

- A. C. Salgado; B. F. Lóscio. Integração de Dados na Web. Centro de Informática - Universidade Federal de Pernambuco, v. 6, p. 157–174, 2001.
- P. Ziegler, K. Dittrich. Three decades of data integration-All problems solved? IFIP congress topical sessions, p. 3–12, 2004.
- Michael J. Franklin, Alon Y. Halevy, and David Maier. From databases to dataspaces: a new abstraction for information management.ACM SIGMOD Rec., 34 (4): 27–33, 2005. DOI:10.1145/1107499.1107502.35

Obrigado!

Dúvidas?