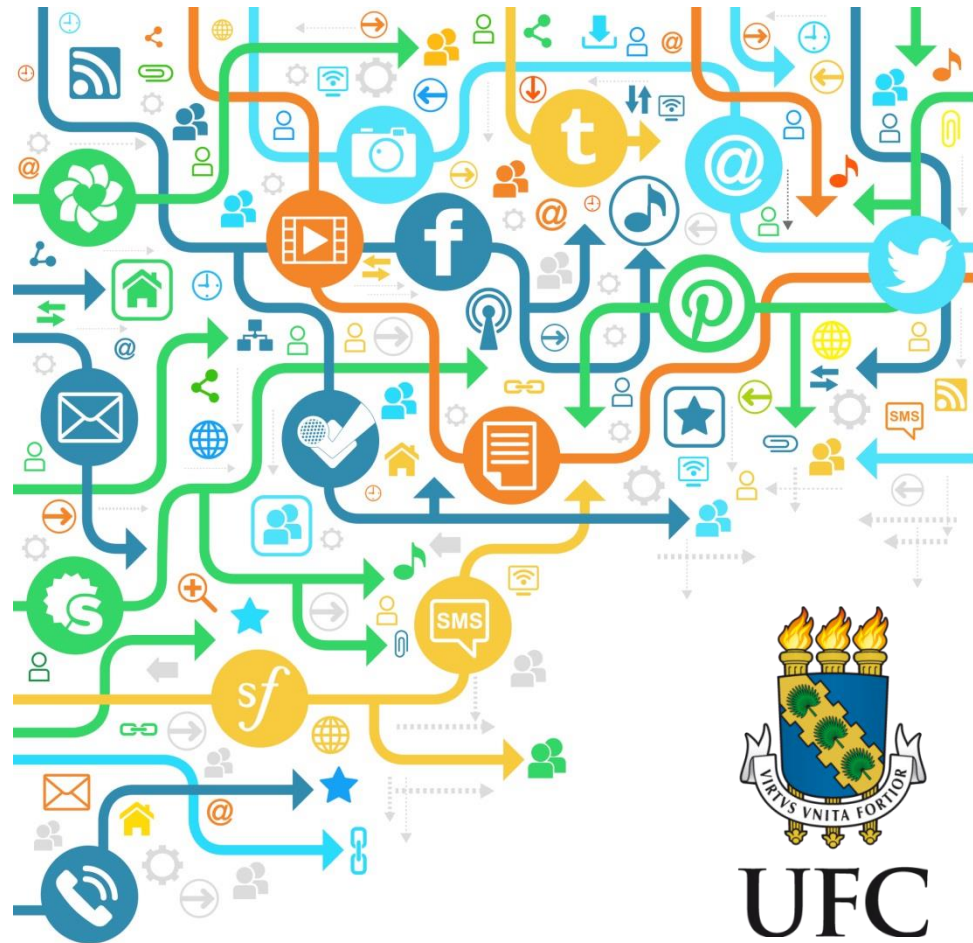


Big Data: Transformando Dados em Riqueza

Regis Pires Magalhães
regismagalhaes@ufc.br
@regispires



As 4 ondas

- Mobilidade
- Social
- Cloud
- **Big Data**

*BIG
DATA*

- **Revolução na maneira de como se adquire e usa tecnologia.**

Big Data?



big data

Web

Imagens

Mapas

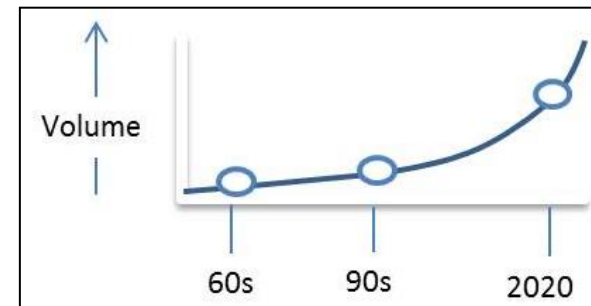
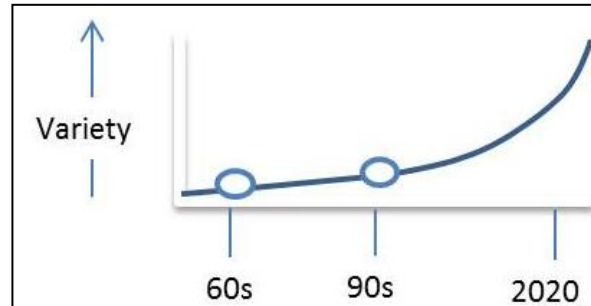
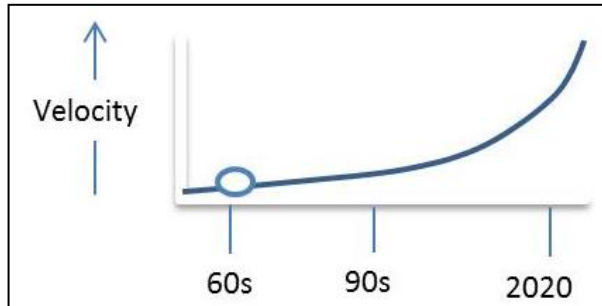
Shopping

Vídeos

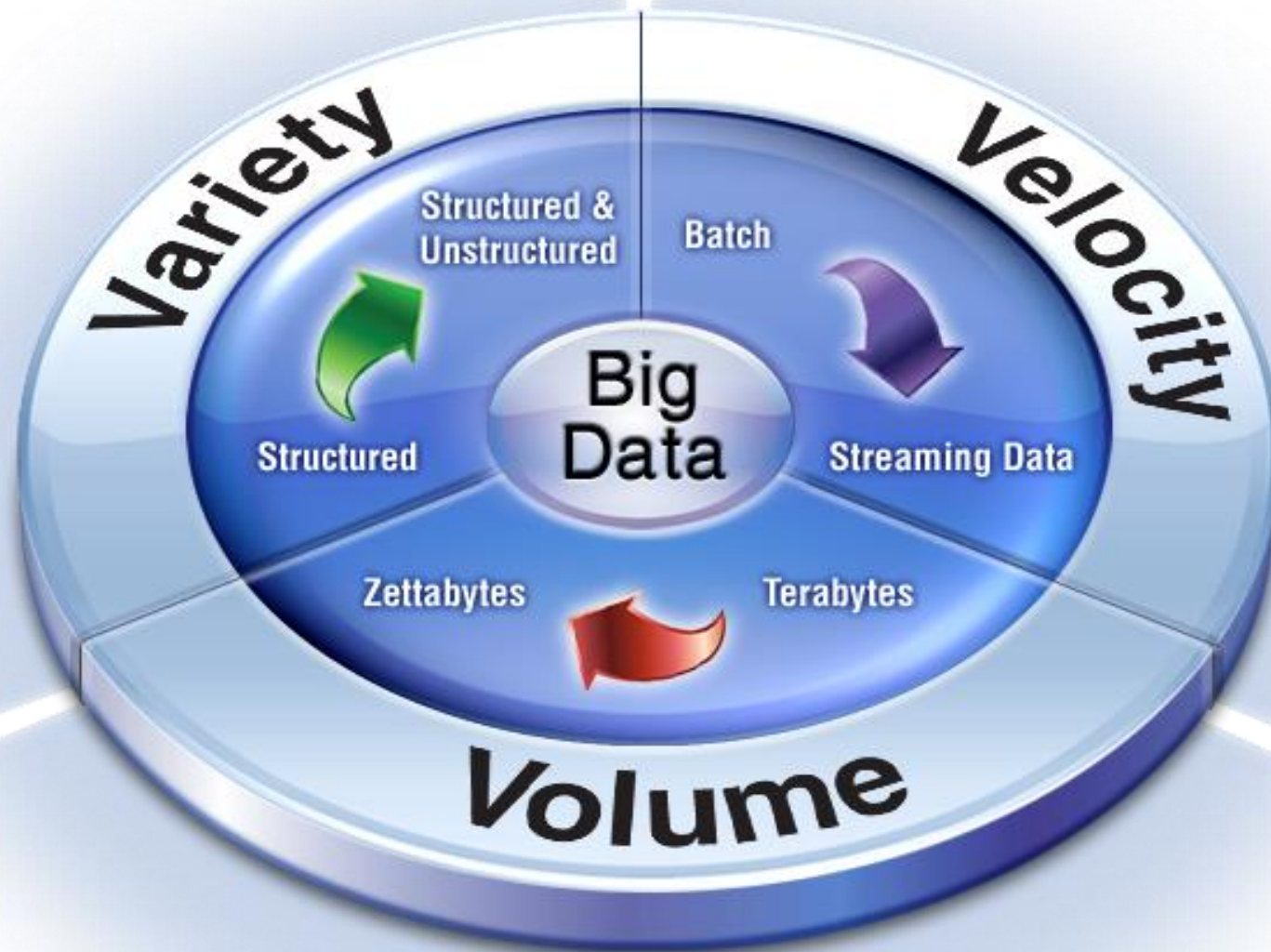
Aproximadamente 1.700.000.000 resultados (0,24 segundos)

O que é Big Data?

- Nova maneira de explorar o imenso volume de dados que circula dentro e fora das empresas.
- Big Data = Grande oportunidade
- **Big Data = Volume + Variedade + Velocidade** (3 Vs) em relação aos dados.



Os 3 Vs de Big Data



Volume

- Geramos petabytes de dados por dia.
- Este volume dobra a cada 18 meses.
- Reaprendendo a contar...
 - Byte, KByte, ...

1,000,000,000,000,000,000,000

Zettabyte Exabyte Petabyte Terabyte Gigabyte Megabyte Kilobyte Byte

Equivalent to:

- Every person on earth tweeting for 100 years
- 125 million years of your favorite 1-hour TV show

Múltiplos do byte					
Prefixo binário (IEC)			Prefixo do SI		
Nome	Símbolo	Múltiplo	Nome	Símbolo	Múltiplo
byte	B	2^0	byte	B	10^0
kibibyte	KiB	2^{10}	Kilobyte	kB	10^3
mebibyte	MiB	2^{20}	megabyte	MB	10^6
gibibyte	GiB	2^{30}	gigabyte	GB	10^9
tebibyte	TiB	2^{40}	terabyte	TB	10^{12}
pebibyte	PiB	2^{50}	petabyte	PB	10^{15}
exbibyte	EiB	2^{60}	exabyte	EB	10^{18}
zebibyte	ZiB	2^{70}	zettabyte	ZB	10^{21}
yobibyte	YiB	2^{80}	yottabyte	YB	10^{24}

Para evitar confusão, foi criada uma nova nomenclatura para diferenciar valores em base 10 e os em base 2.

World's Data Volume

800 Terabytes, 2000

160 Exabytes, 2006

500 Exabytes, 2009

2.7 Zettabytes, 2012

35 Zettabytes by 2020



Big Data is growing fast

Annual growth rate

60%

Structured and unstructured data¹

In social media alone,
every 60 seconds

600



The digital universe
will grow to

2.7zB

in 2012, up

Petabytes

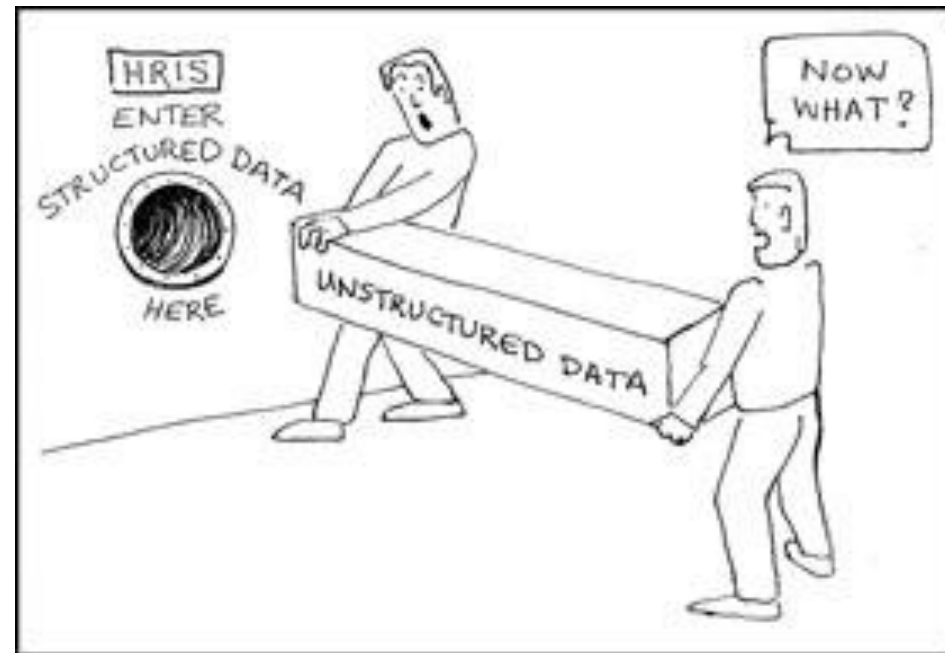


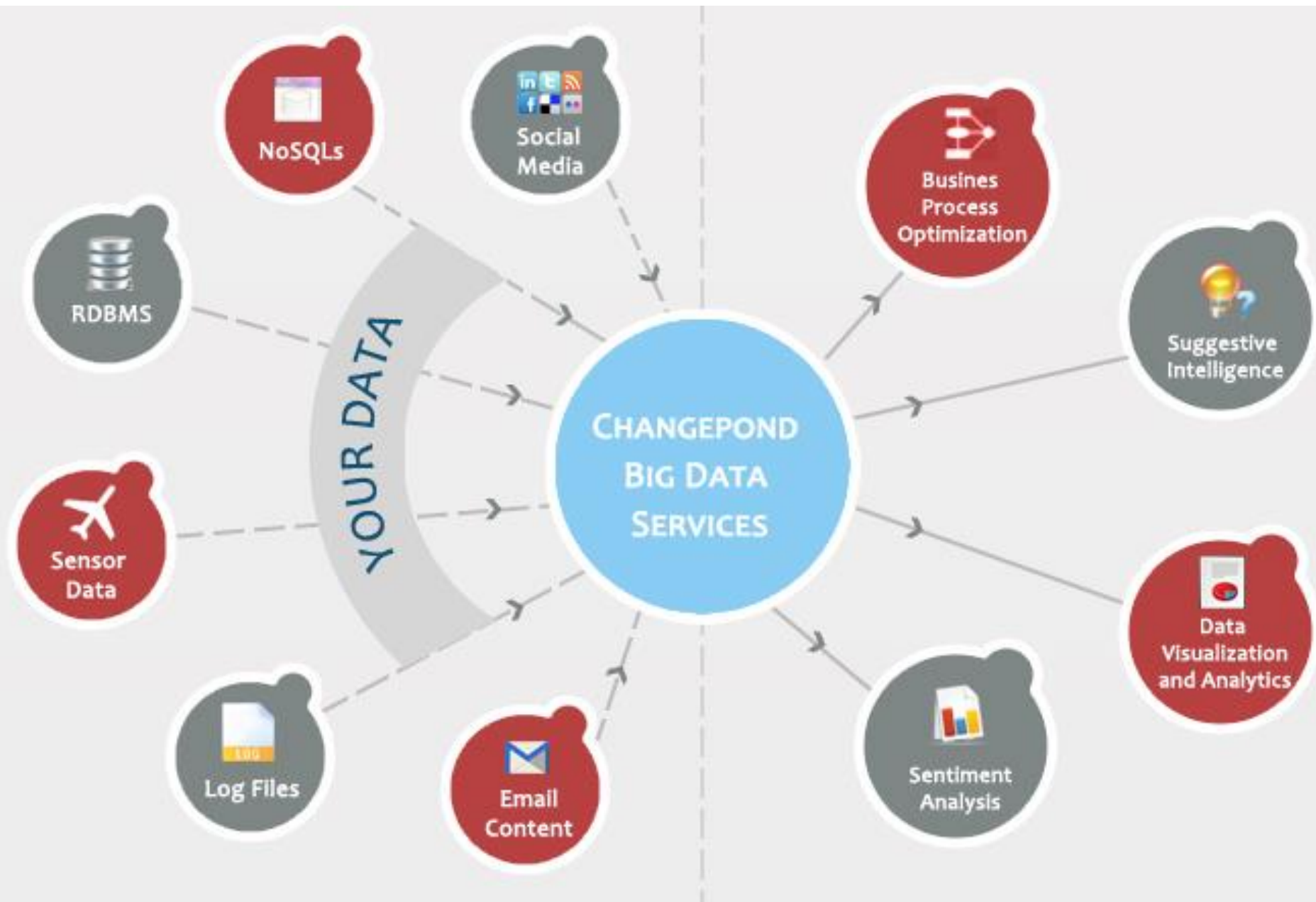
Tudo já escrito pela humanidade em toda a história: **50 petabytes**.

O Google processa **20 petabytes** por dia.

Variedade

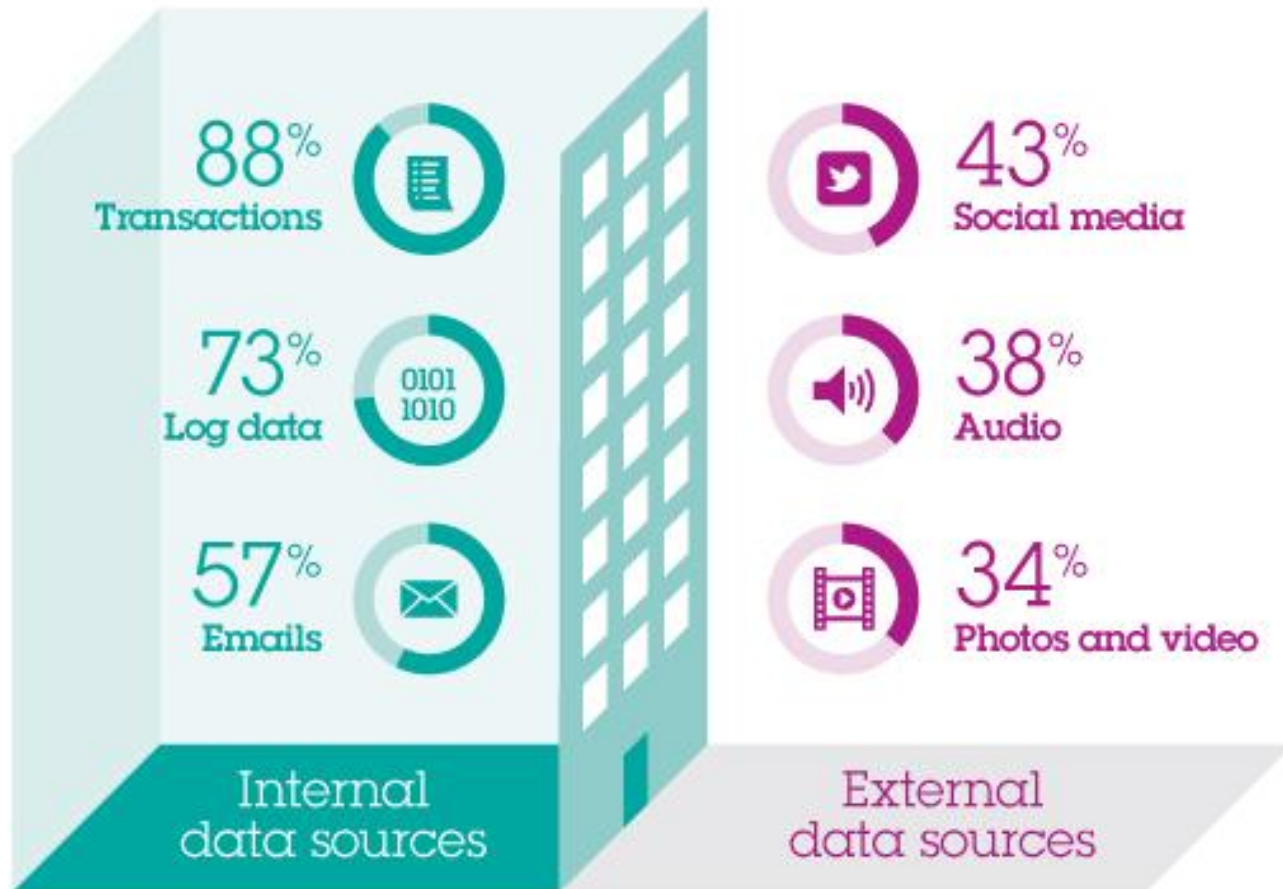
- Dados de:
 - Sistemas **estruturados** (minoria);
 - Sistemas **não estruturados** (imensa maioria).
 - Gerados por e-mail, mídias sociais (Facebook, Twitter, YouTube e outros).
 - Documentos eletrônicos (textos, apresentações, planilhas, etc).
 - Mensagens instantâneas.
 - Sensores.
 - Etiquetas RFID.
 - Câmeras de vídeo.
 - Etc.





Where does big data come from?

Most big data efforts are currently focused on analyzing internal data to extract insights. Fewer organizations are looking at data outside their firewalls, such as social media.

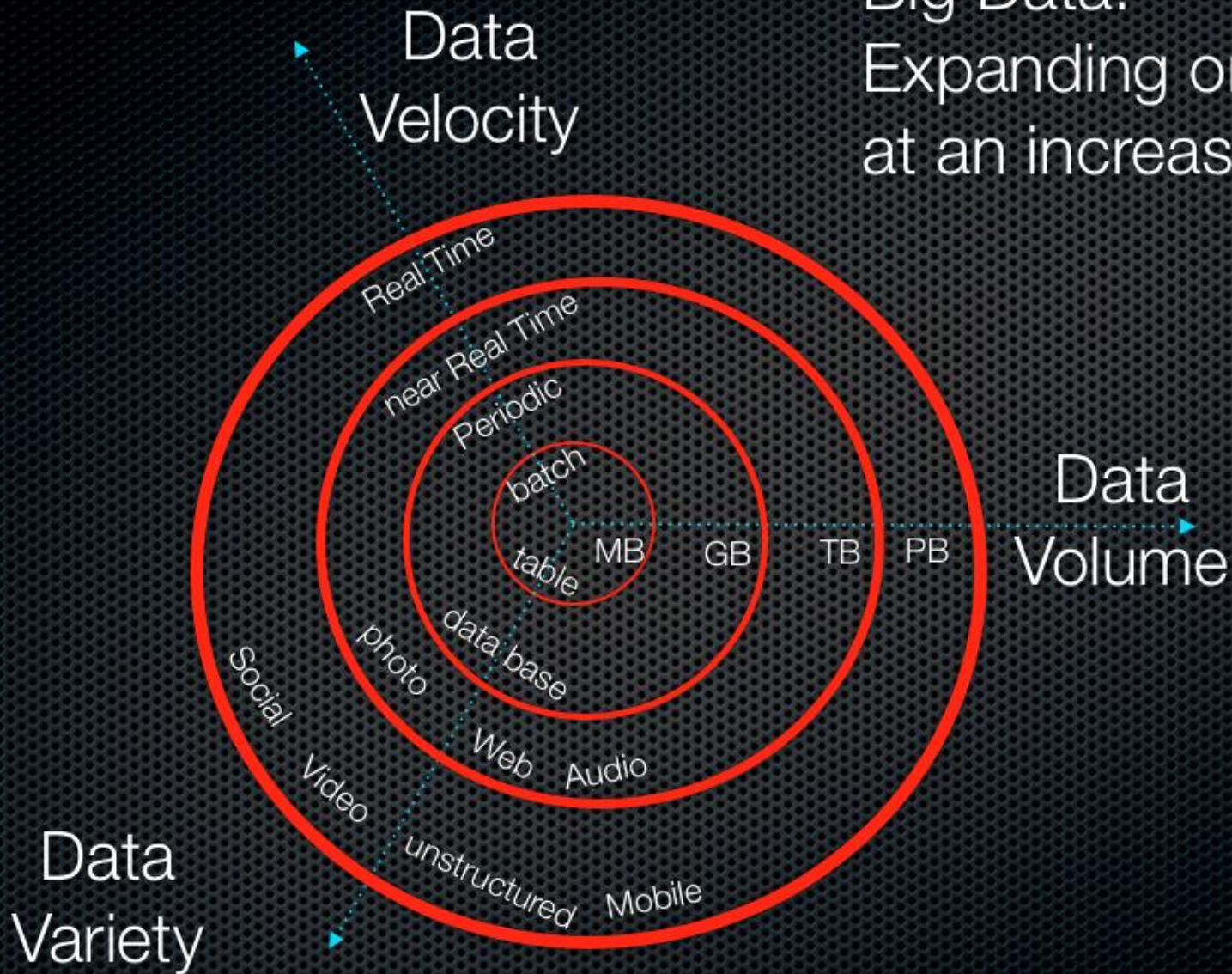


Velocidade

- Muitas vezes precisamos agir praticamente em **tempo real** sobre esse imenso volume de dados.
 - Ex: controle automático de tráfego nas ruas.
- **Stream processing** permite tratamento de dados em tempo real.
 - Mineração de dados em tempo real.
 - Corrente contínua de dados (streaming data) atravessando um conjunto de queries.
 - Exemplo:
 - Detecção de falhas em fábrica de semicondutores através da monitoração em tempo real do processo de detecção e classificação de falhas.



Big Data:
Expanding on 3 fronts
at an increasing rate.



Acrescentando mais 2 Vs...

Big Data = **Volume** + **Variedade** +
Velocidade + **Veracidade** + **Valor**



Veracidade

- Os dados devem fazer sentido, ser autênticos e ter consistência.



Valor

- Armazenamento, uso e análise dos dados devem gerar valor (retorno) sobre esses investimentos realizados.

- Exemplo:

- Na área de seguros, a análise de fraudes pode ser melhorada, minimizando os riscos, utilizando, por exemplo, a análise de dados que circulam nas mídias sociais.



Valor



- Só faz sentido se o valor da análise dos dados compensar o custo de sua coleta, armazenamento e processamento.
- Existem também questões legais a serem resolvidas.



Big Data como ferramenta

“Big Data é uma ferramenta que pode ser usada em diversos nichos.”

“A maneira de usar esses dados, de correlacionar as informações é que faz diferença.”

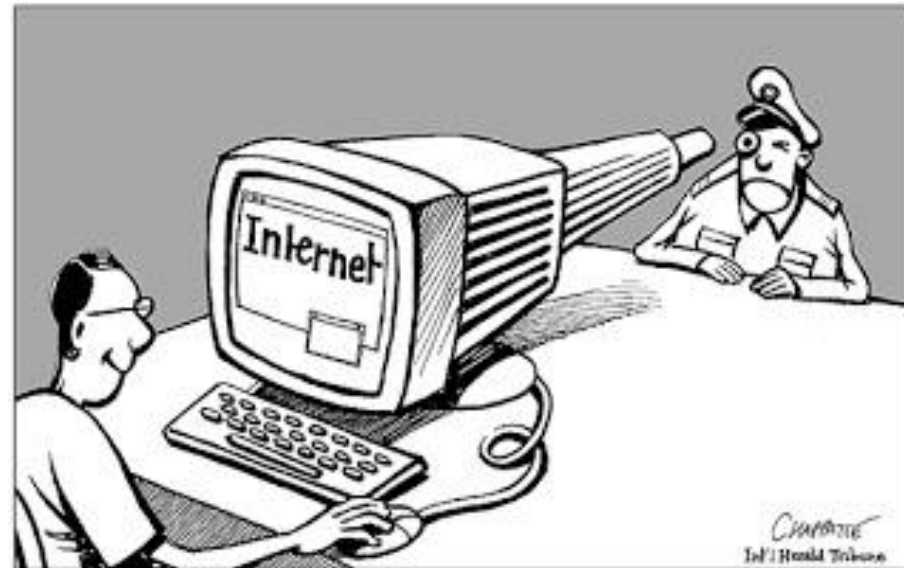
Ricardo Souza

Consultor de projetos da Oracle



Segurança e privacidade

- Temos também a questão da privacidade e acesso a dados confidenciais.
- É essencial criar uma política de acesso e divulgação das informações.

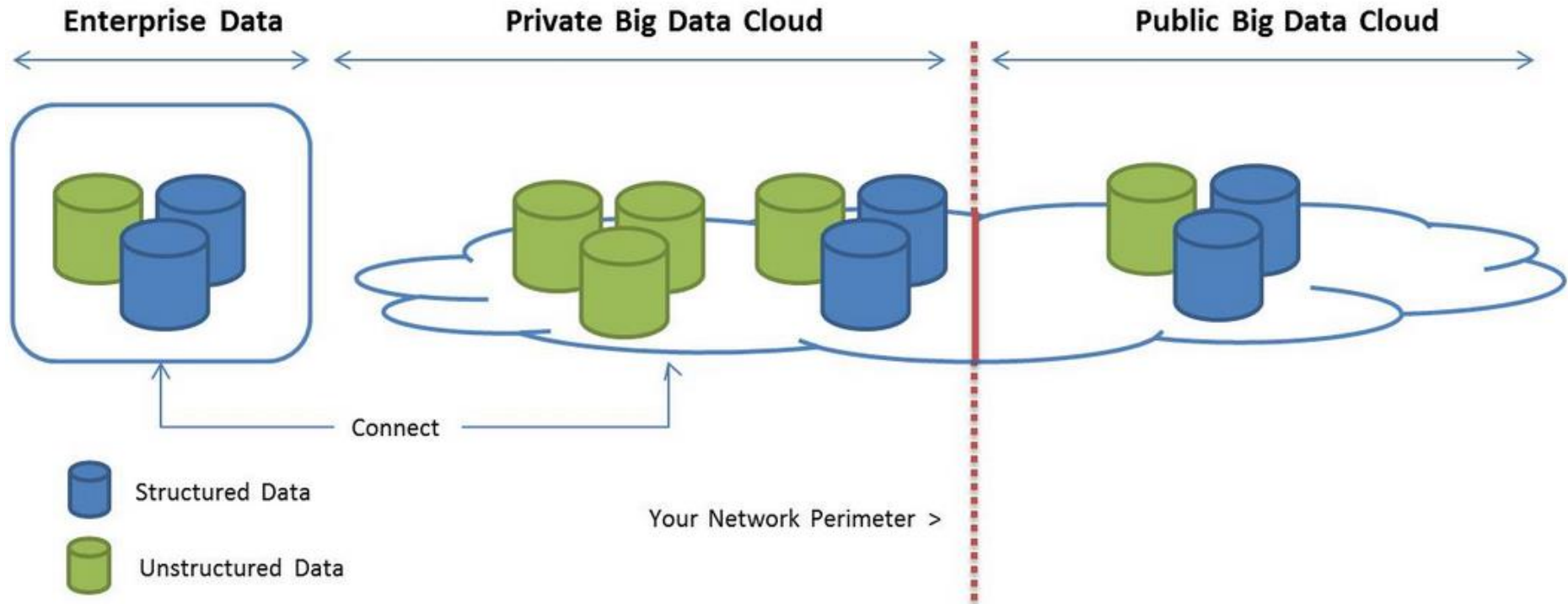


Computação em Nuvem x Big Data

- Computação em Nuvem é um impulsionador de Big Data.
 - Uso de nuvens públicas para suportar imensos volumes de dados.



Big Data Cloud



Credit: Watalon.com : Big Data Cloud

Todos, inclusive grandes empresas usando intensamente as redes sociais

- Pode-se rapidamente extrair informações extremamente importantes a partir disso, evitar grandes prejuízos e aperfeiçoar bastante os serviços e produtos.



Potencial de Big Data

Big data tem o potencial de se tornar a próxima fronteira para inovações, competição e lucro.



Benefícios de Big Data


Não basta ter os dados.

Usar os dados para transformá-los em lucro é o grande desafio que alguns já estão concretizando.



HOW DO YOU EXPECT TO BENEFIT BY USING BIG DATA?

COST SAVINGS IN BUSINESS PROCESSES  61%

COST SAVINGS IN IT  57%

COMPETITIVE ADVANTAGE  35%

INCREASED PROFITS FROM THE BUSINESS MODEL  35%

DON'T KNOW  3%

Aplicações de Big Data

Smarter Healthcare



Multi-channel



Finance



Log Analysis



Homeland Security



Traffic Control



Telecom



Search Quality



Manufacturing



Trading Analytics



Fraud and Risk



Retail: Churn, NBO



Setores com maior potencial de benefício

- Saúde
- Governo
- Comércio
- Indústria
- Tecnologia de localização pessoal



Saúde



- Em 10 anos o uso de Big Data na saúde capturou \$300 bilhões anualmente.
- Diminuição dos gastos governamentais em 8%.
- Áreas
 - **P&D**: Pesquisa e desenvolvimento; projeto de testes clínicos; medicina personalizada.
 - **Clínica**: transparência em dados clínicos e suporte à decisão clínica.
 - **Contabilidade**: Detecção avançada de fraudes; preços de medicamentos baseado no desempenho.
 - **Saúde pública**: fiscalização da saúde pública; sistemas de resposta.
 - **Novo modelo de negócios**: Agregação de registros de pacientes, plataformas online e comunidades.

Governo

- Aumento da produtividade / eficiência através do meio digital.
- O uso de Big Data no setor público Europeu reduziu custos em 20% ou 300 bilhões de Euros.



Governo



- Em 01/Set/2013, o Serviço Federal de Processamento de Dados (Serpro) vai lançar o primeiro serviço de computação em nuvem do governo federal.
- O ambiente vai abrigar sistemas para o **Programa Cidades Digitais**, e serão oferecidas soluções de **educação, atendimento médico hospitalar, gestão e comunicações** para cerca de 200 **municípios brasileiros**.

Comércio / Serviços

- Aumento de lucratividade e produtividade.
- Aumento potencial de margens operacionais: 60%.
- Sistemas de recomendação.

Indústria

- Aumento da produção e diminuição dos custos.
- Potencial de diminuir os custos operacionais em 50% em todos os setores industriais.



Tecnologia de Localização Pessoal

- O volume de dados de localização pessoal aumentou rapidamente com a adoção de smartphones.
- O potencial é imenso, pois não está em um único setor, mas distribuído por todos eles.
- Áreas
 - **GPS**: navegação incluindo roteamento inteligente baseado em tráfego em tempo-real.
 - **Marketing**: propaganda baseada na localização.
 - **Social**: rastreamento de pessoas, compartilhamento de localização, entretenimento.

Gestão de Pessoas

- Melhoria contínua do ambiente de trabalho a partir de ações que meçam a satisfação do funcionário.
- Identificar habilidades que precisam ser mais bem desenvolvidas entre os colaboradores para se alinhar às necessidades ou tendências de mercado.
- Seleção de funcionários com perfis específicos.
- Dentro das empresas, os funcionários deixam pegadas digitais.
- Ciência da força de trabalho
 - Objetivo: avanços no relacionamento entre empresas e funcionários.

Educação

- **MOOCs**

- Massive Open Online Courses
 - EdX, Coursera, Udacity, etc.

- **Learning analytics**

- Ferramenta usada para decifrar tendências e padrões a partir de big data disponível sobre o aprendizado dos alunos.
- Recurso útil para fazer escolhas pedagógicas a partir da necessidade dos alunos.
- Tornar o processo de orientação dos estudantes muito mais preciso.

Bancos de Dados Relacionais

- Modelo relacional proposto por Edgar F. Codd, pesquisador da IBM, em 1969.
- Demanda:
 - Acessar dados estruturados gerados pelos sistemas internos das corporações.
- Não adequado para:
 - Dados não estruturados.
 - Volumes na casa dos petabytes de dados.
- Para tratar dados na escala de volume, variedade e velocidade **precisamos de outros modelos.**

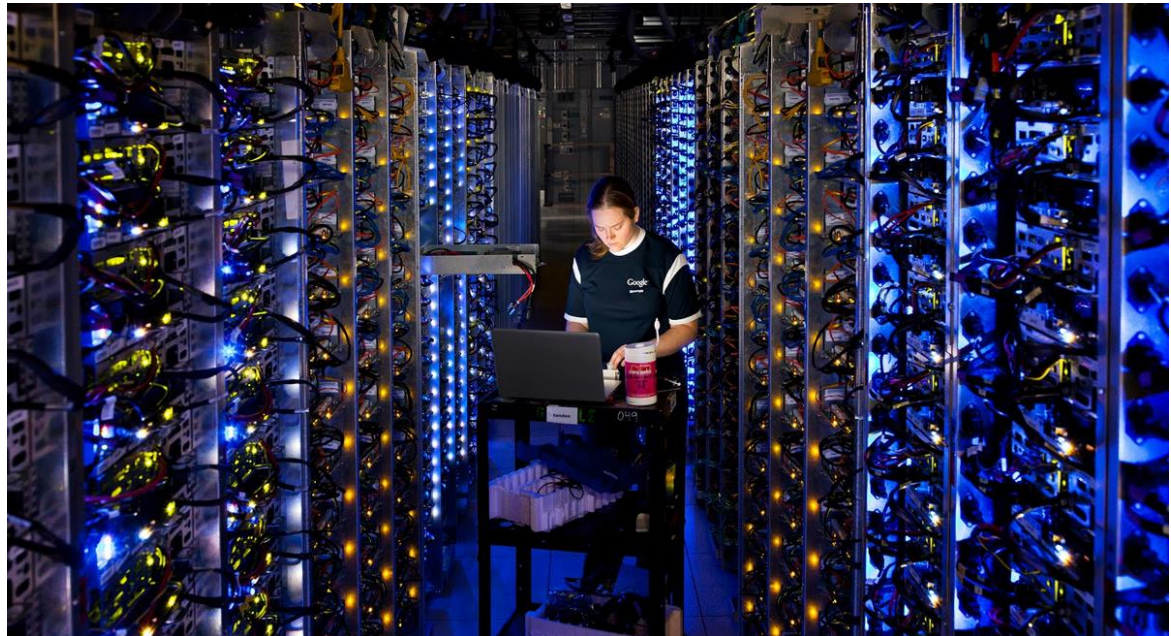
Tecnologias

- **Infraestrutura**

- Armazenam e processam os petabytes de dados.
- Tecnologias:
 - Bancos de Dados **NoSQL (Not Only SQL)**

- **Analytics**

- Tecnologias:
 - Hadoop e MapReduce.



Hadoop



- **Hadoop =**

- Sistema de Arquivos Distribuído (HDFS) +
- Framework de Processamento (MapReduce).
 - Escrever MapReduce não é tarefa simples.
- **HDFS** → simples (sem recursos avançados).



- **HBase**

- Modelado a partir do Google BigTable.

- **Hive** (criado pelo Facebook)

- Criado pelo Facebook.
- Datawarehouse em um cluster Hadoop.
- Interface semelhante a SQL.
- Permite criação de tabelas.

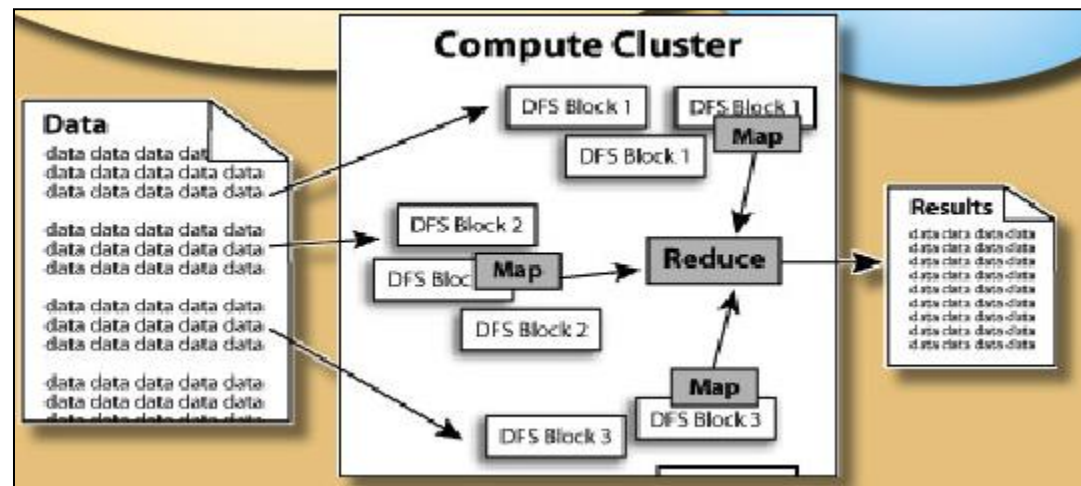
- **Pig** (criado pelo Yahoo!)

- Linguagem de fluxo de dados (dataflow) para processar grandes volumes de dados de forma fácil e rápida.
 - Facilita a escrita de MapReduce.

- Consultas Hive e Pig são transformadas em Jobs MapReduce.

MapReduce

- Os dados são distribuídos entre os computadores de um cluster.
- Programas **Map** em cada computador analisam e processam seu subconjunto dos dados e retornam resultados intermediários como pares chave-valor.
- O passo **Reduce** ordena e combina os resultados intermediários para retornar um resultado final.



Bancos de dados NoSQL

- Capazes de tratar imensos volumes de dados estruturados e não estruturados.
- Tipos / Orientados a:

- **Colunas / Tabular**



Cassandra

APACHE
HBASE

- **Google Big Table** – usado internamente pelo Google e Google App Engine.
- Apache **HBase** (baseado no Big Table); Apache **Cassandra** (baseado no DynamoDB da Amazon).
 - Facebook (criador do Cassandra) substituiu Cassandra por HBase (Mensagens).
 - Netflix usa Cassandra como BD de seus serviços de streaming.
 - Twitter usa Cassandra (Analytics, TopTweets, ...) e MySQL (para Tweets)

- **Documento**

- **MongoDB**, Apache **CouchDB** (documentos JSON).



- **Chave/Valor**

- **DynamoDB** da Amazon; Riak; **Redis**; Cache; Voldemort.

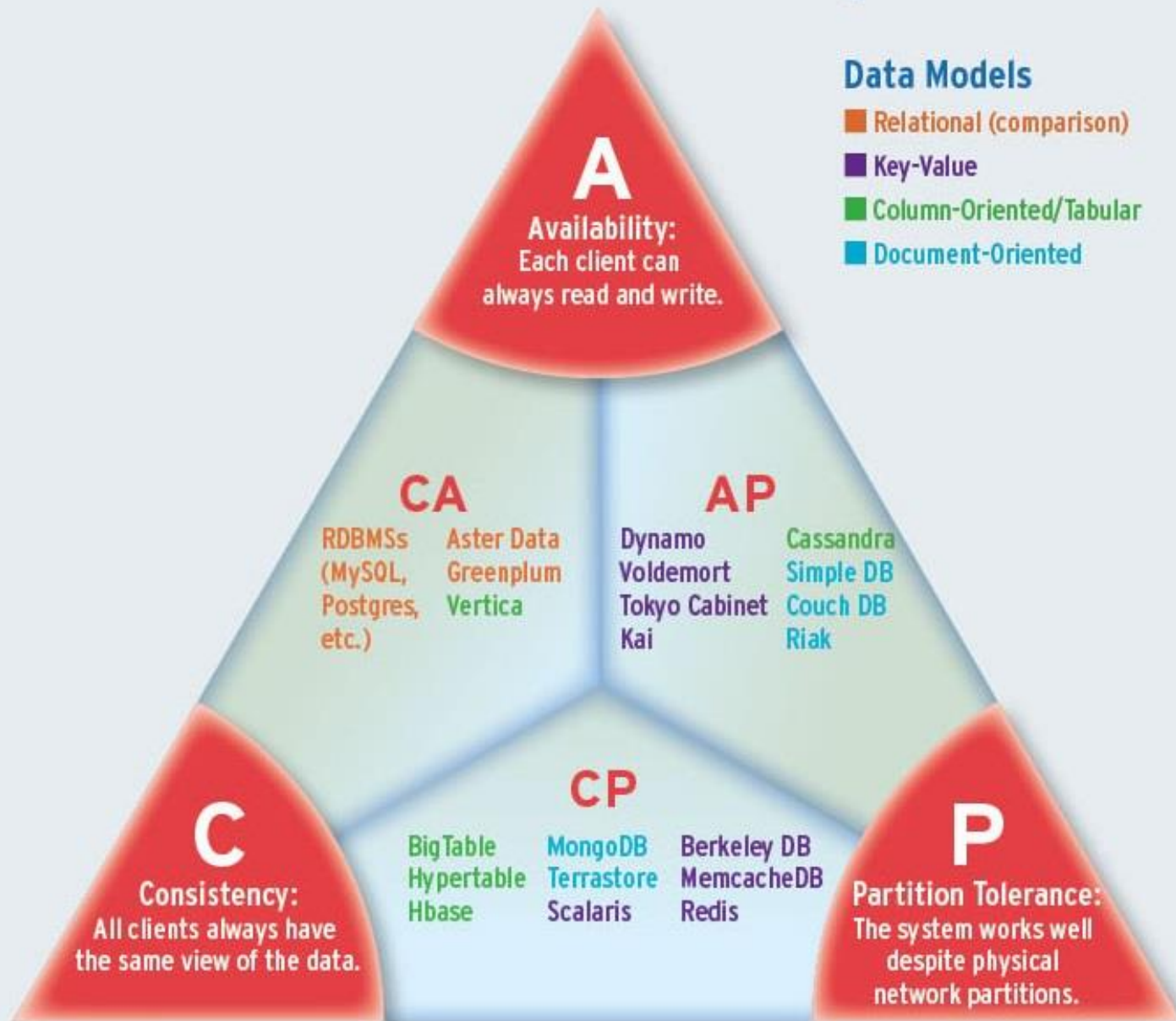


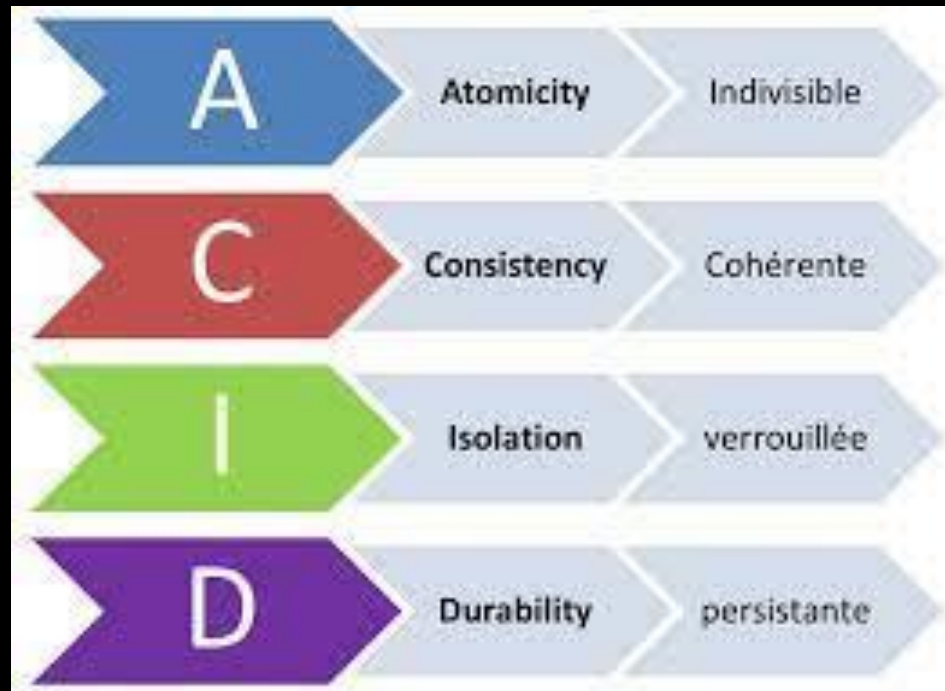
- **Grafo**

- **Neo4j**, Allegro, Virtuoso.



Visual Guide to NoSQL Systems





Examples



ORACLE

- ✓ ACID transactions
- ✓ SQL support
- ✓ Standardized
- ✗ Horizontal Scaling
- ✗ High Availability

RDBMS (SQL)

Examples



cassandra

redis

- ✓ Horizontal Scaling
- ✓ High Availability
- ✗ ACID transactions
- ✗ SQL support
- ✗ Standardized

NoSQL

Examples



VOLTDB

Cloud Spanner



CockroachDB

- ✓ ACID transactions
- ✓ Horizontal Scaling
- ✓ High Availability
- ✓ SQL support
- ✗ Standardized

NewSQL

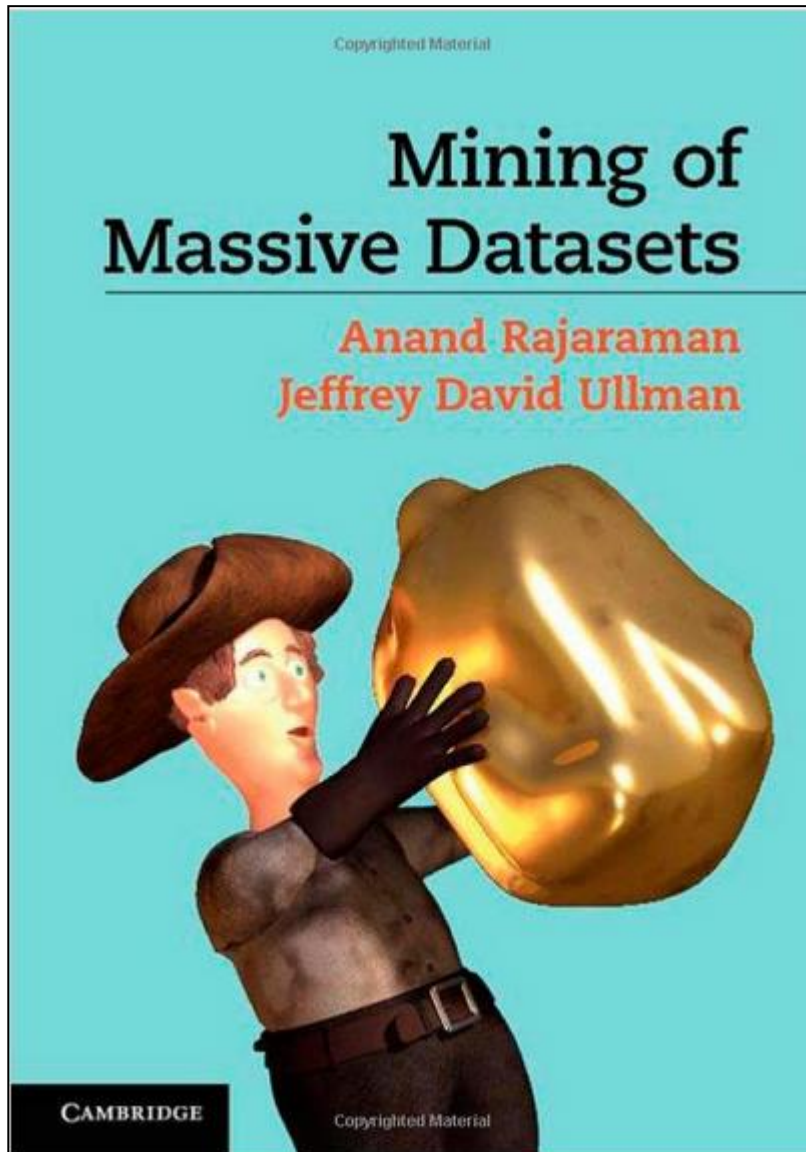


Parameters for Comparison	RDBMS	NoSQL	NewSQL
SCHEMA	Relational Schema / TABLE	Schema-free	BOTH
SCALABILITY	Scalable reads	Scalable writes/reads Hirizontal Scalable	Scalable writes/reads Hirizontal Scalable
High Availability	Custom High-Availability	Auto High-Availability	Built-in High-Availability
STORAGE	On-Disk + Cache	On-Disk + Cache	On-Disk + Cache
Cloud Support	Not Fully	SUPPORTED	FULLY SUPPORTED
Query Complexity	LOW	High	Very High
ACID-CAP-BASE	ACID	CAP Through BASE	ACID
OLTP	Not Fully Supported	Not Supported	Fully Supported
Performance Overhead	Huge	Moderate	Minimal
Security Concerns	Very Very High	LOW	LOW
Examples	Oracle, MS SQL, MySQL, IBM DB2, PostgreSQL	MongoDB, Cassandra, Redis, HBase	Google Spanner, VoltDB
Use Cases	Financial , CRM, HR Applications	Big Data, IoT, Social Network Applications	Gaming, E-Commerce (High Availability), Telecom industry

Table 1: Comparison between SQL, NoSQL and NewSQL

Illustration is by Dr. Rabi Prasad Padhy, PhD

Mining of Massive Datasets



Jure Leskovec, Stanford University
Anand Rajaraman, WalmartLabs
Jeffrey David Ullman, Stanford University

<http://www.mmds.org/>

- Algoritmos Práticos de mineração de dados sobre Big Data.
- Map-reduce → ferramenta para paralelizar algoritmos.
- Tratamento de dados sensíveis à localização.
- Processamento de streaming de dados.
- PageRank.
- Clustering.
- Descoberta de Itens Frequentes.
- Sistemas de Recomendação.
- Propagandas na Web.
- Large-Scale Machine Learning
- Neural Nets and Deep Learning

Designing Data-Intensive Applications:

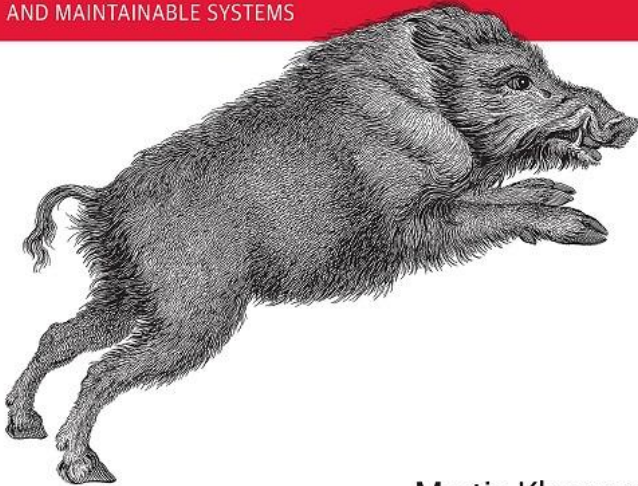
The Big Ideas Behind Reliable, Scalable, and Maintainable Systems

por Martin Kleppmann

O'REILLY®

Designing Data-Intensive Applications

THE BIG IDEAS BEHIND RELIABLE, SCALABLE,
AND MAINTAINABLE SYSTEMS



Martin Kleppmann

I. Foundations of Data Systems

1. Reliable, Scalable, and Maintainable Applications
2. Data Models and Query Languages
3. Storage and Retrieval
4. Encoding and Evolution

II. Distributed Data

5. Replication
6. Partitioning
7. Transactions
8. The Trouble with Distributed Systems
9. Consistency and Consensus

III. Derived Data

10. Batch Processing
11. Stream Processing
12. The Future of Data Systems

Spring Data com tecnologias NoSQL

- <https://spring.io/projects/spring-data>
- Spring Data dá suporte a:
 - MongoDB
 - Neo4J
 - Elasticsearch
 - Solr
 - Redis
 - Cassandra
 - Couchbase
 - LDAP
 - Gemfire
- **Spring Boot** provê configurações automáticas para as tecnologias acima.
 - <https://docs.spring.io/spring-boot/docs/current/reference/html/boot-features-nosql.html>

Hibernate OGM

<http://hibernate.org/ogm/>

- Object Grid Mapper.
- API Java completa para armazenar dados em NoSQL stores.
- Interface comum (JPA) para várias abordagens NoSQL.
 - Usa JPQL como linguagem de consulta.
 - Suporte para: MongoDB (documento), Ehcache (chave-valor), Infinispan (chave-valor), Neo4j (grafo).
 - Em andamento: Cassandra, Redis, CouchDB.
- CRUD para entidades JPA.

Conclusões

- Big Data é real e está sendo cada vez mais usado nos mais diversos segmentos.
- Big Data abre muitas oportunidades de pesquisa e desenvolvimento de software.
- A onda de Big Data tem um vínculo muito forte com as outras 3 ondas: computação em nuvem, mobilidade e redes sociais.
- É preciso aliar TI + Negócio.
- Algumas tecnologias promissoras: **MapReduce** (Hadoop), **Bancos NoSQL** (MongoDB, Cassandra, HBase, Redis, etc.)

Referências

- Você realmente sabe o que é Big Data?
 - https://www.ibm.com/developerworks/mydeveloperworks/blogs/ctaurion/entry/voce_realmente_sabe_o_que_e_big_data?lang=en
- A quantas anda o Big Data no atual mercado de tecnologia?
 - <http://imasters.com.br/banco-de-dados/a-quantas-anda-o-big-data-no-atual-mercado-de-tecnologia/>
- Big data é um tsunami em alto mar – Resenha do livro Big Data
 - <http://imasters.com.br/tecnologia/redes-e-servidores/resenha-do-livro-big-data/>
- Serviço em nuvem do governo federal será lançado em setembro
 - <http://imasters.com.br/noticia/servico-em-nuvem-do-governo-federal-sera-lancado-em-setembro/>
- Conhecendo o Hadoop
 - <http://www.bigdatabrasil.net/conhecendo-o-hadoop/>
- Considerações sobre o Banco de Dados Apache Cassandra
 - <http://www.bigdatabrasil.net/consideracoes-sobre-o-banco-de-dados-apache-cassandra/>
- <http://hibernate.org/ogm/>
- <https://docs.spring.io/spring-boot/docs/current/reference/html/boot-features-nosql.html>

Obrigado!

Dúvidas, comentários,
sugestões?

big data

Regis Pires Magalhães
regismagalhaes@ufc.br
@regispires