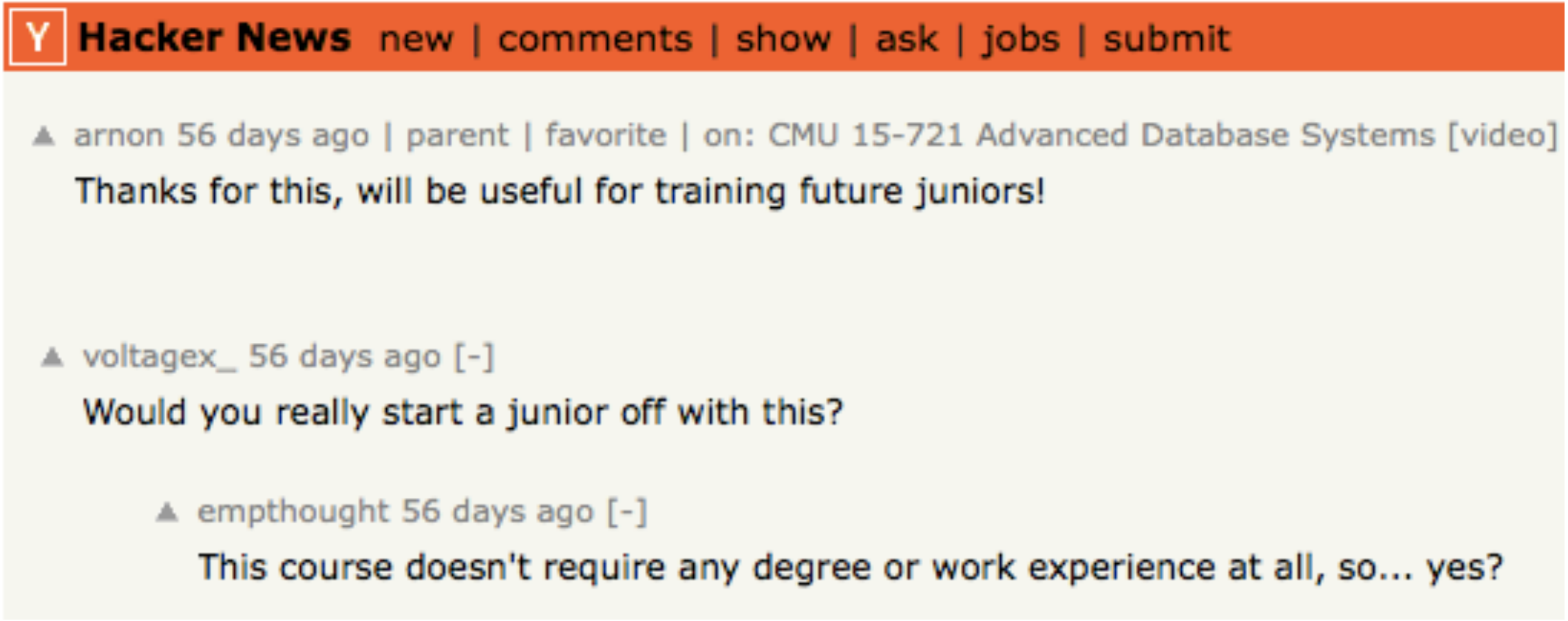
BigData

A Bullet for The General (1967)

Aula #1 - Processamento de BigData

EDUARDO CUNHA DE ALMEIDA

Preciso de formação em Computação?



Visão do curso

Este curso discute os diversos sistemas de processamento de Big Data sem aprofundar em conceitos básicos da computação.

NÃO substitui cursos de algoritmos
e estruturas de dados ou sistemas
de bancos de dados.

-> veja **www.inf.ufpr.br** (Bacharelado em CC)

Agenda do Curso

- Modelos de Armazenamento
- Modelos de Indexação
- Modelagem de dados: abstrato e lógico
- Processamento de consultas (Simone)
- Processamento distribuído de dados (Ramiro)

Horário

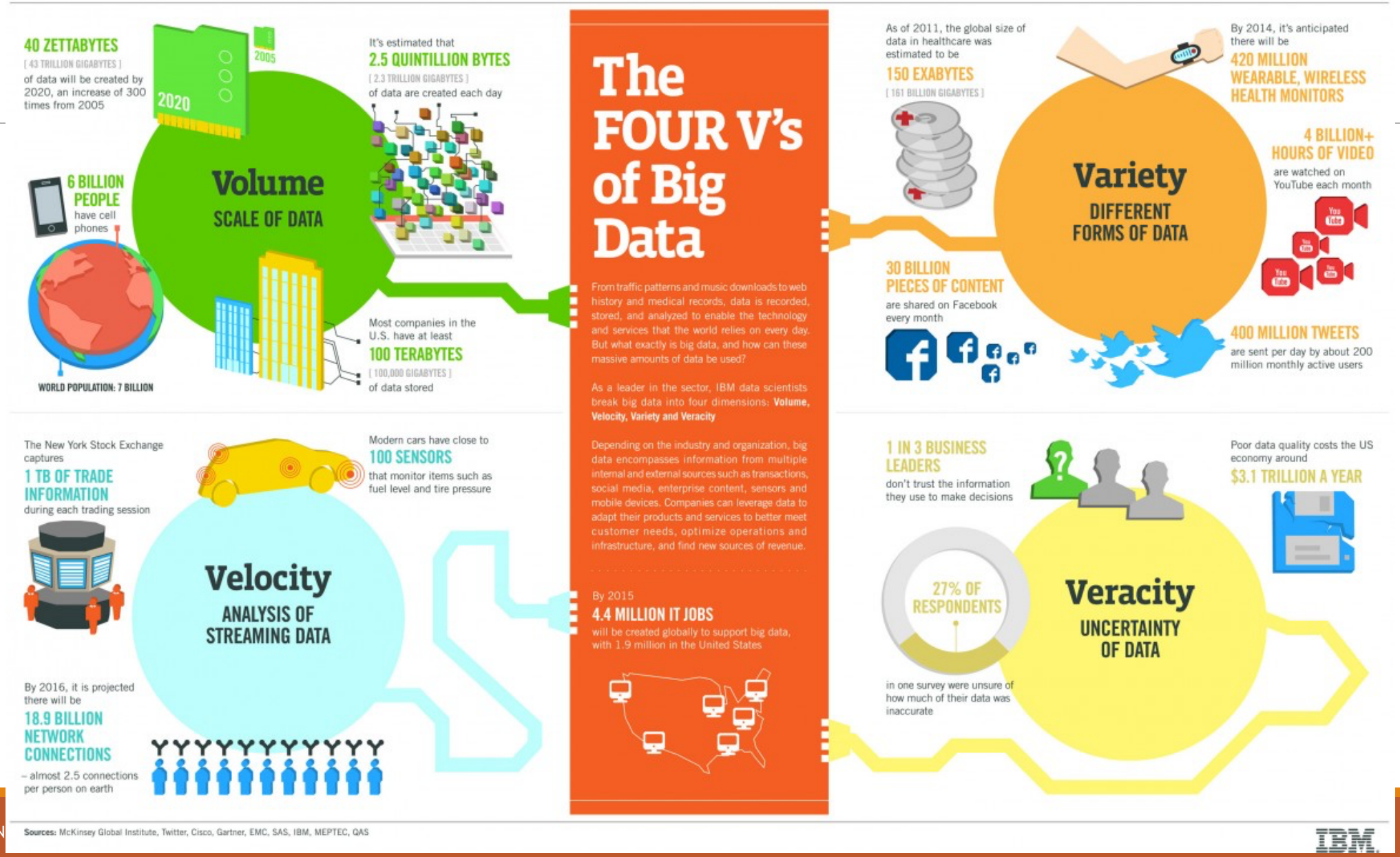
- Aulas começam hoje
- 04/08 - 14/09: Eduardo
- 22/09 - 20/10: Simone
- 27/10 - 30/11: Ramiro

Uma Definição de BigData

"Big Data é qualquer dado que seja computacionalmente caro de gerenciar e difícil de extrair valor."

Michael Franklin, UC Berkeley

Outra Definição de BigData



Desafios e Oportunidades

O dado é o novo petróleo

"Data is the new Oil"
Clive Humby, CNBC

"We want to enable people to leverage Big Data by developing systems and platforms that are reusable and scalable across multiple application domains."

bigdata@CSAIL
MIT BIG DATA INITIATIVE

"... information systems are moving from the back office, to being the backbone of business value creation."
Ed Dumbill, Forbes

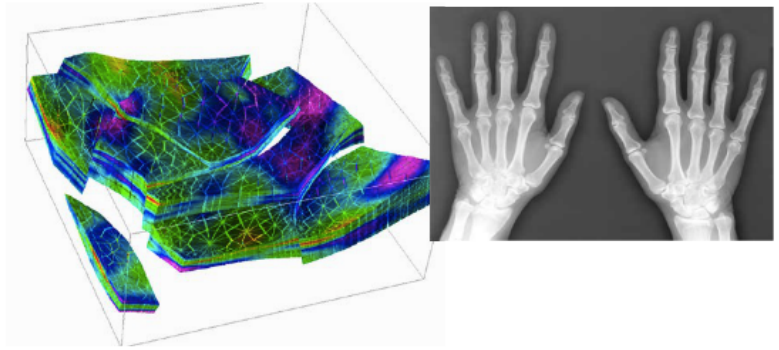
Forbes

"Hadoop Vendor MapR Brings in $110 Million, Led by Google Capital"
Steven Norton, The Wall Street Journal

THE WALL STREET JOURNAL.

Fontes de Big Data

- Áudio
- Transações
- Videos
- Texto
- Estatísticas
- Imagens



… gerando dados em 2014 …

- Large Array Telescope:
20,000 PB/Day in 2020 according to IBM

- 15TB/Day

- 1800 trans/sec

- 60,000 tweets/sec
(World Cup`14)

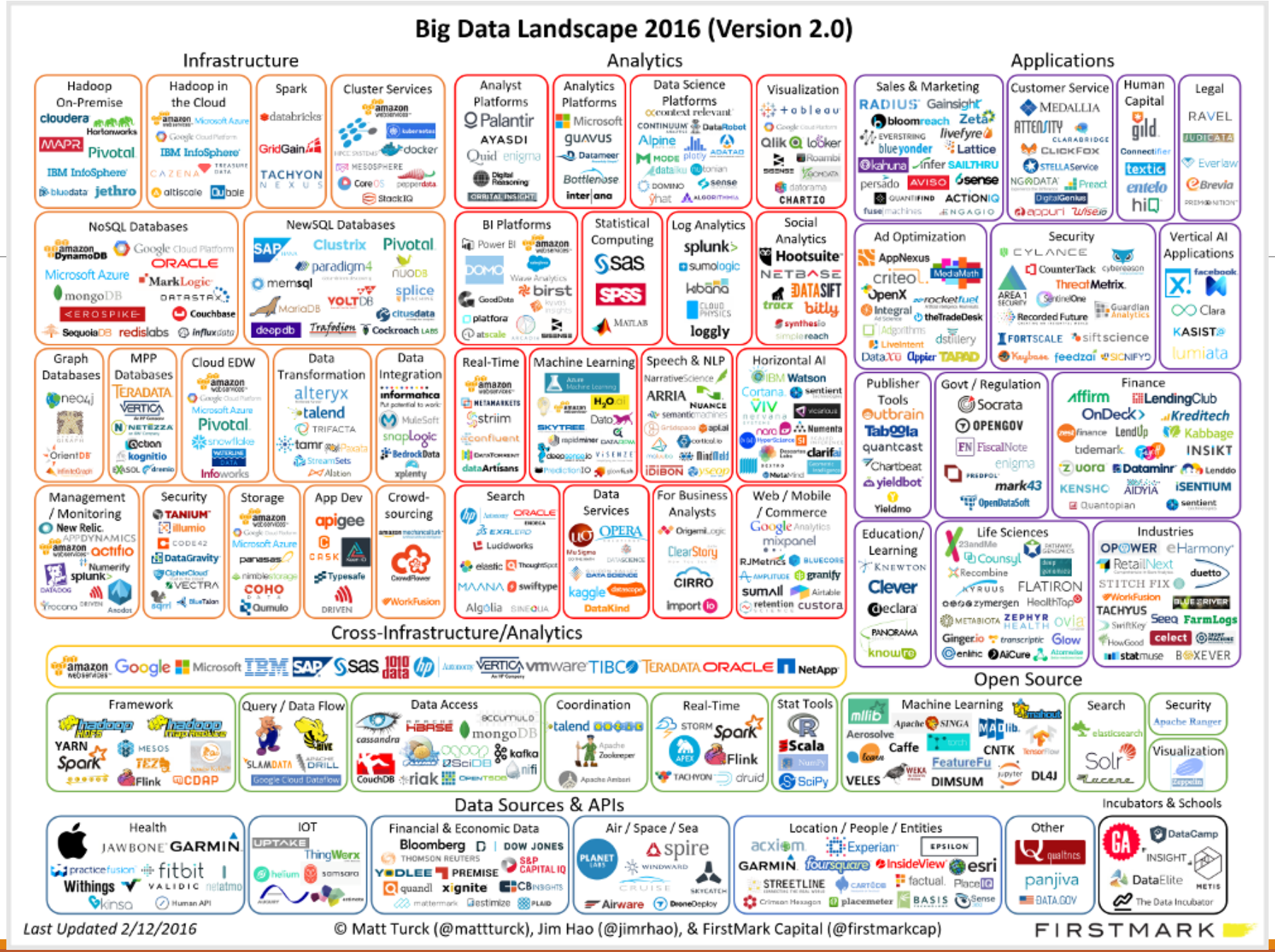
- 1 trillion stored objects

- LHC: 1GB/sec

2017 This Is What Happens In An Internet Minute

2018 This Is What Happens In An Internet Minute

… e podemos comprar também !!


Big Data Landscape 2016 (Version 2.0)
Infrastructure
Analytics
Applications
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
Sales & Marketing
Customer Service
Human Capital
Legal
NoSQL Databases
NewSQL Databases
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Ad Optimization
Security
Vertical AI Applications
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Publisher Tools
Govt / Regulation
Finance
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
Education / Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Incubators & Schools
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other
Last Updated 2/12/2016
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
FIRSTMARK

Soluções de processamento Big data (neste curso)

System-R (“Processamento tradicional”)

- Projeto anos 70: relações and consistência forte de dados

NewSQL

- System-R com esteróides

NoSQL

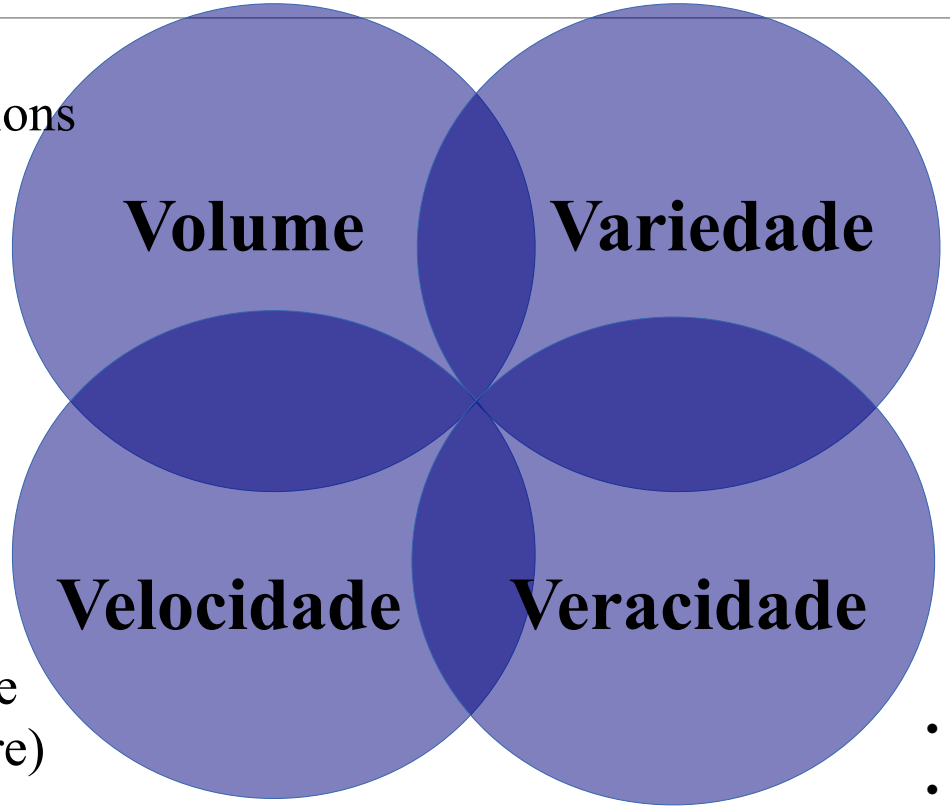
- Consistência eventual

MapReduce

- Modelo de programação de “divisão e conquista”

O que é BigData?

- On-line transactions
- Documentos
- Emails
- Blogs
- Social Media



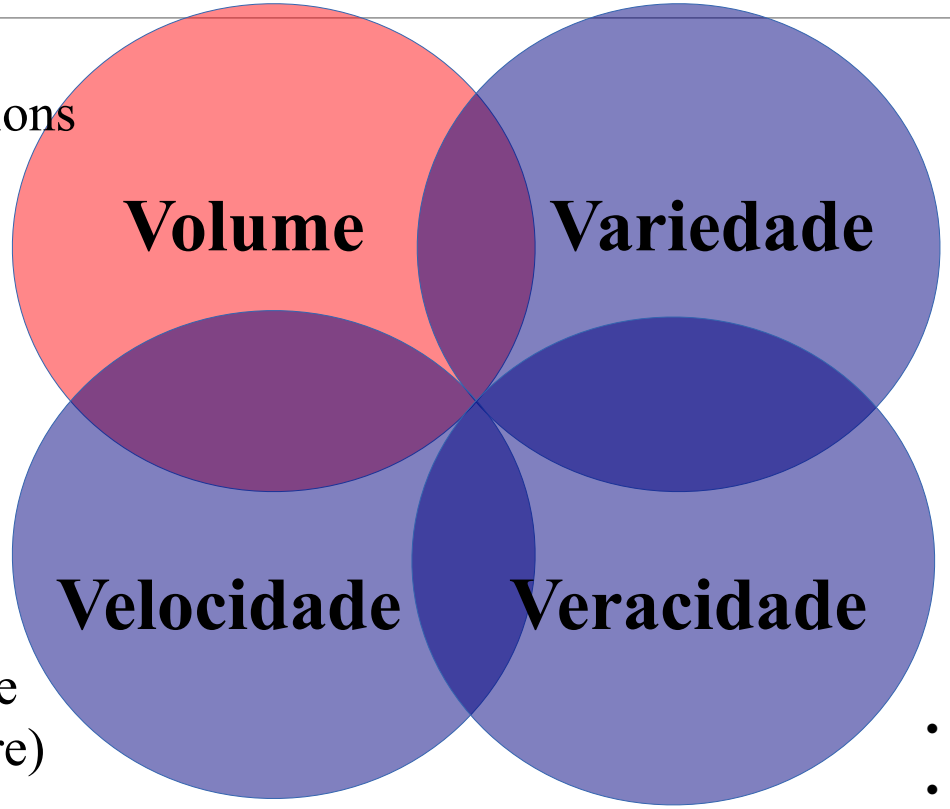
- Estruturado
- Semi-estruturado
- não estruturado

- Transações on-line (anytime/anywhere)
- Streaming

- Structured
- Semi-structured
- Unstructured

O que é BigData?

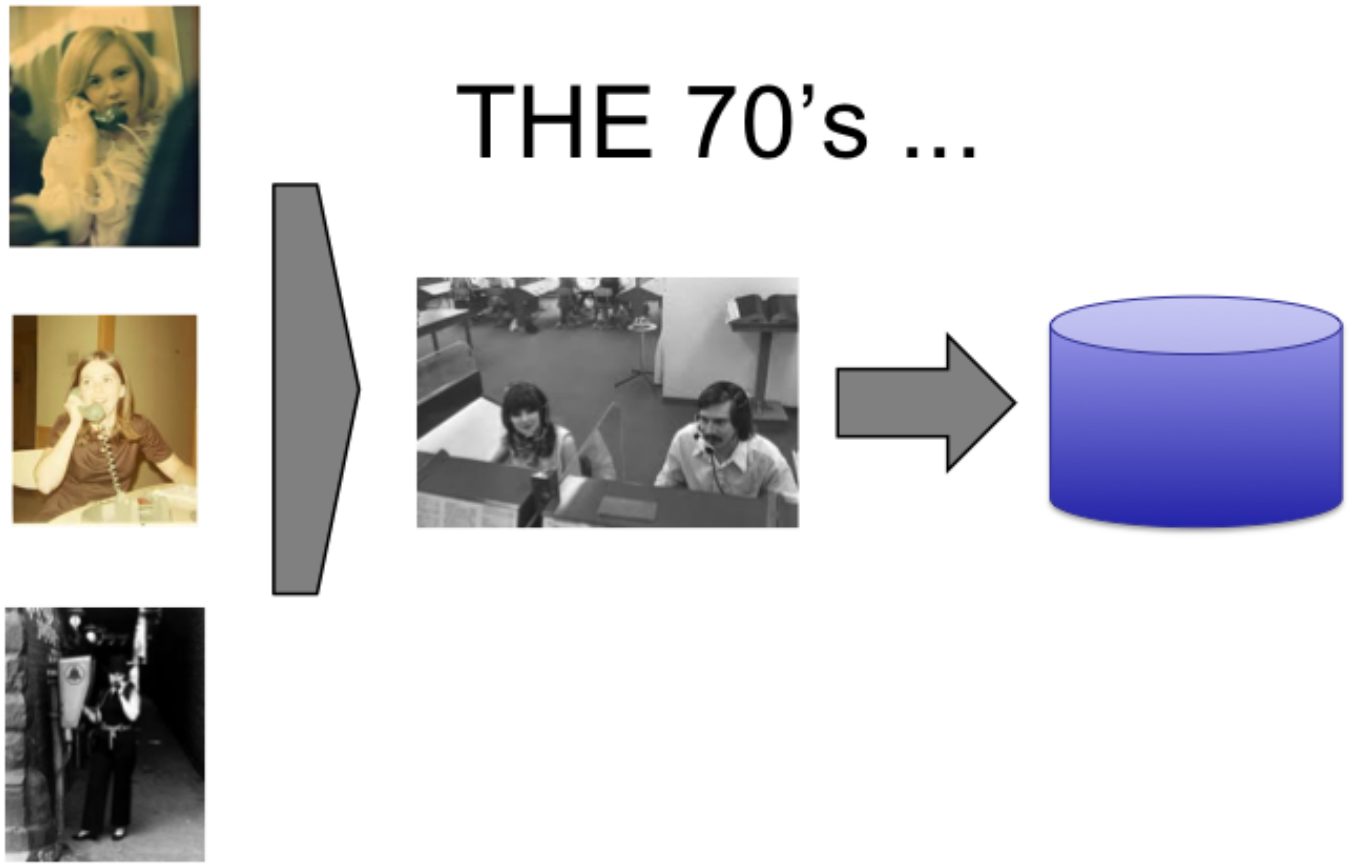
- On-line transactions
- Documentos
- Emails
- Blogs
- Social Media



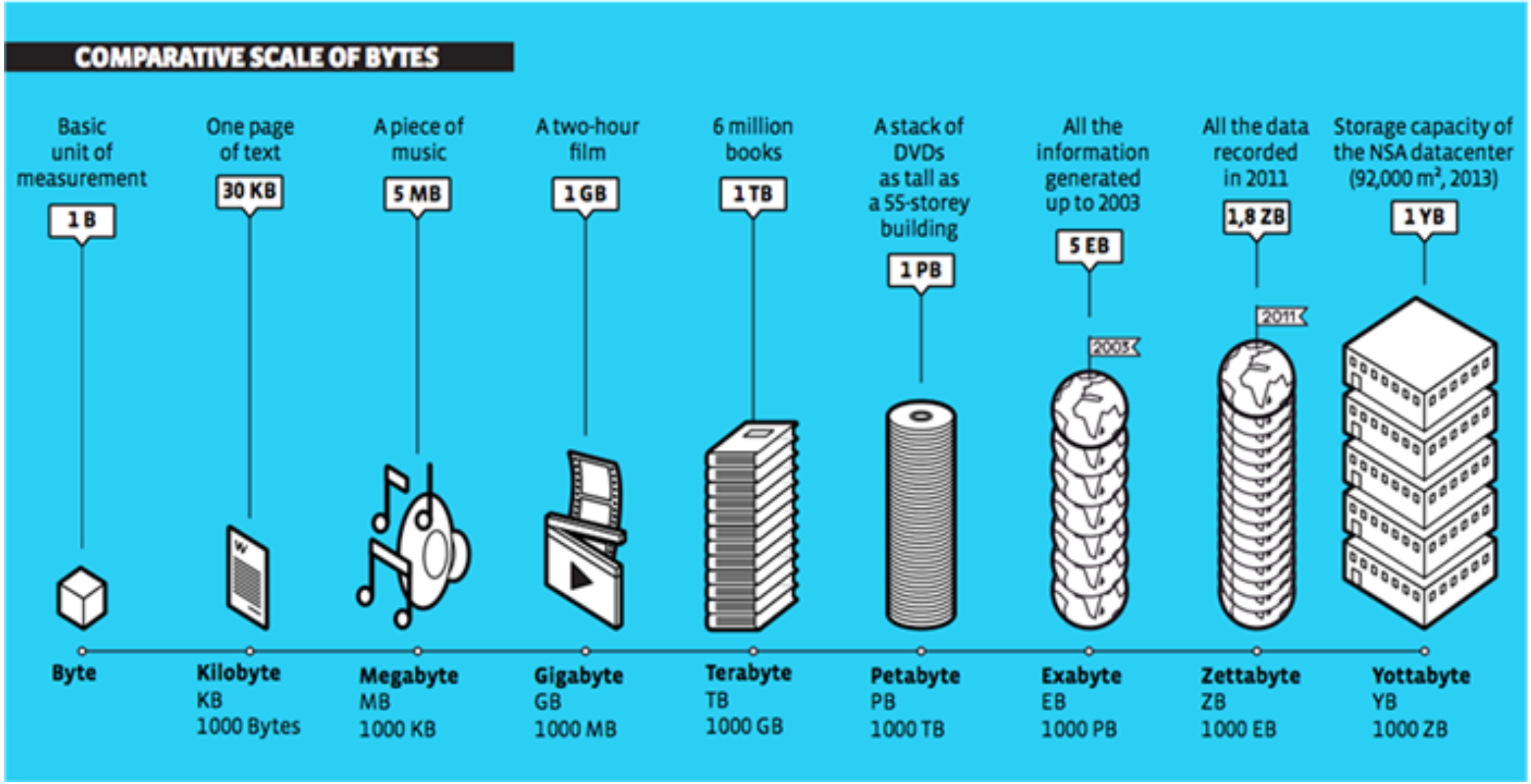
- Estruturado
- Semi-estruturado
- não estruturado

- Transações on-line (anytime/anywhere)
- Streaming

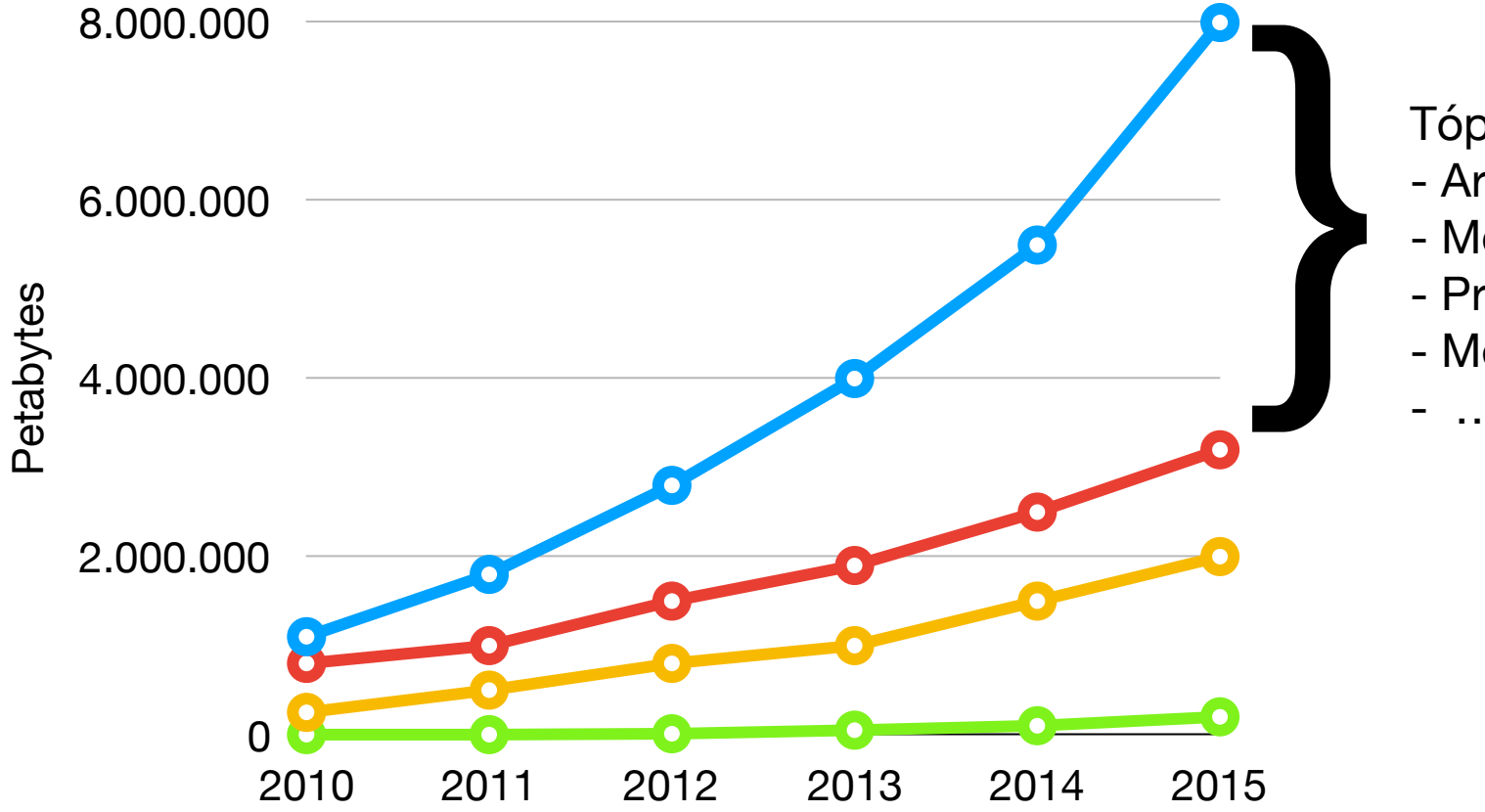
- Structured
- Semi-structured
- Unstructured


THE 70’s ...

* M. Stonebraker et al, 2009. The End of an Architectural Era (It's Time for a Complete Rewrite).



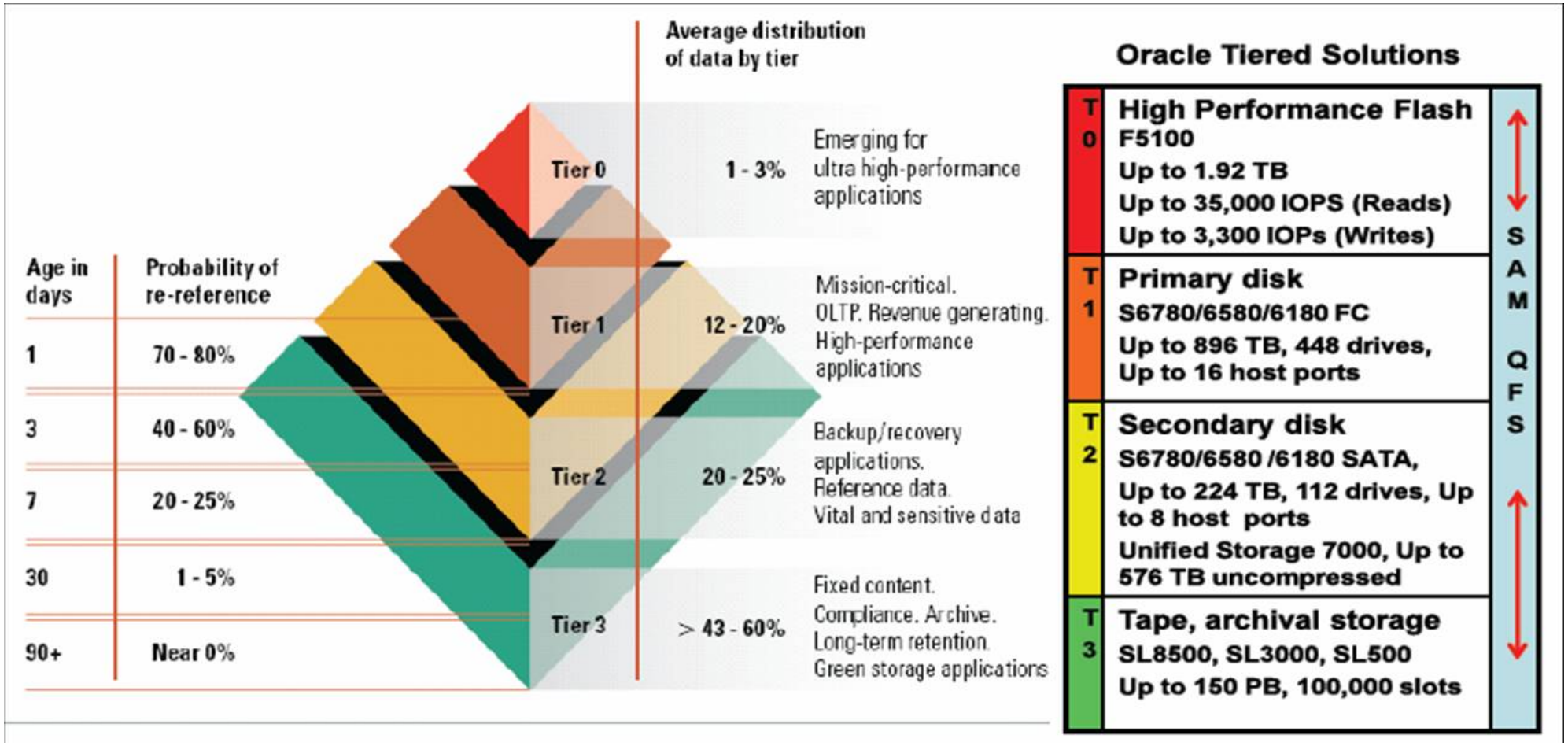
[Patrice Koehl, UC Davis]
https://er.educause.edu



Tópicos:
- Armazenamento
- Modelos de dados
- Processamento
- Métodos de acesso
- ...

Todd Walter, “Big Plateaus of Big Data on the Big Island”, VLDB 2015

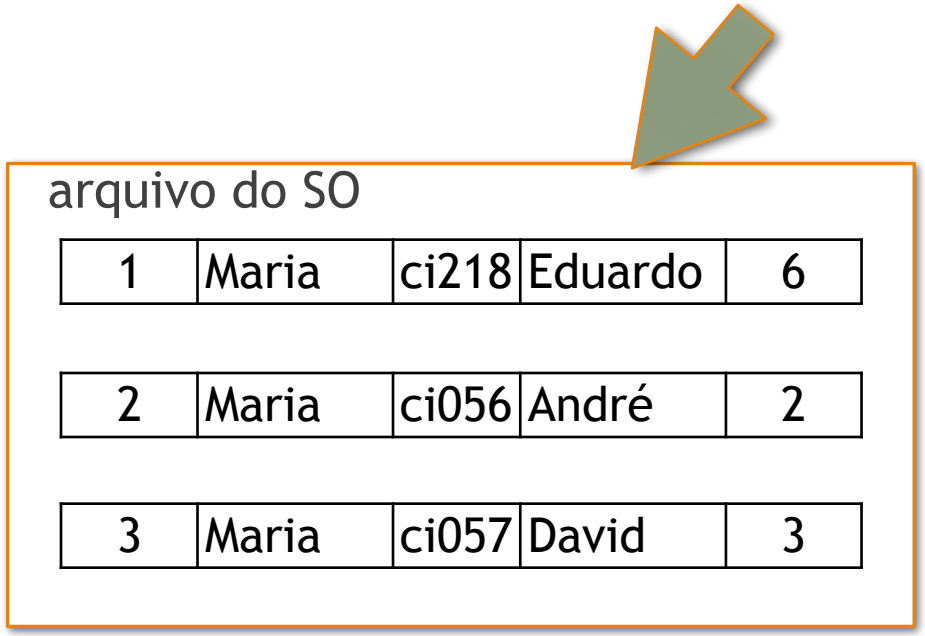
Armazenamento em Camadas



Google's
datacenter

Armazenamento tupla vs. coluna

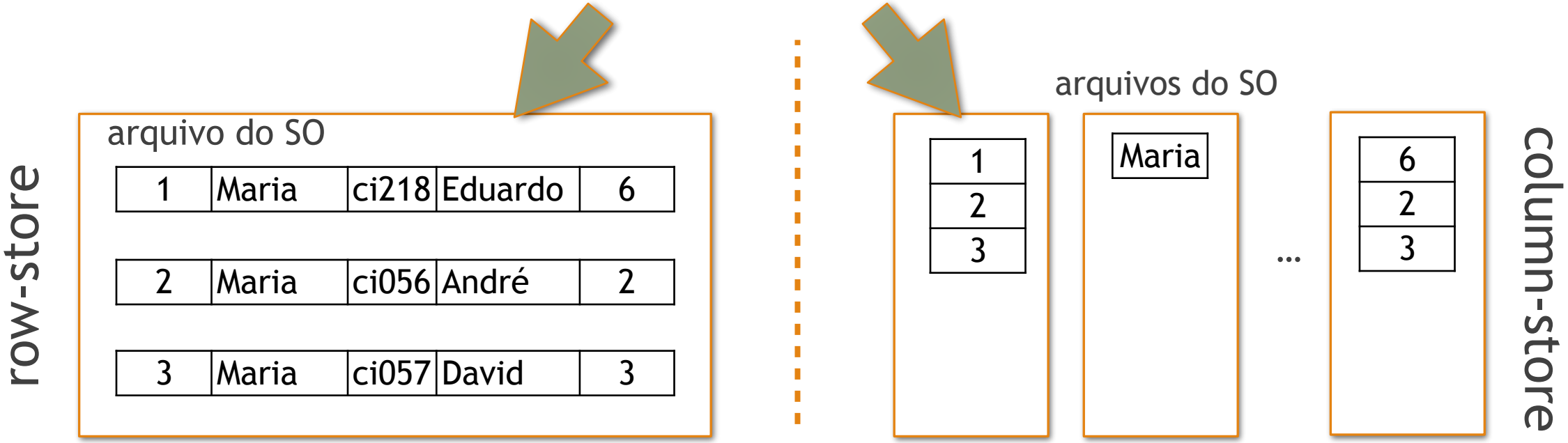
matricula	aluno	disciplina	instrutor	semestre
1	Maria	ci218	Eduardo	6
2	Maria	ci056	André	2
3	Maria	ci057	David	3



row-store

Armazenamento tupla vs. coluna

matricula	aluno	disciplina	instrutor	semestre
1	Maria	ci218	Eduardo	6
2	Maria	ci056	André	2
3	Maria	ci057	David	3



Armazenamento tupla vs. coluna

row-store

- Escritas feitas diretamente num arquivo
- Leituras buscam dados desnecessários

column-store

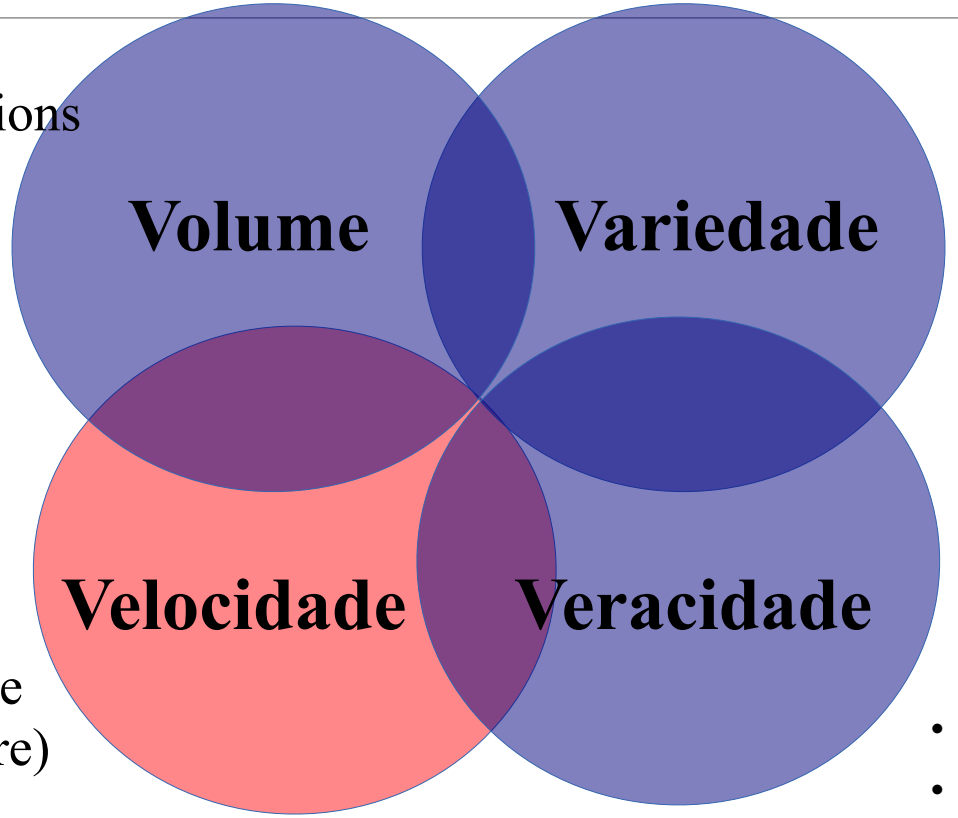
- Leituras somente buscam dados relevantes
- Fácil compressão
- Escritas necessitam de fraccionamento dos dados

Sistemas para Volume

1. C-Store (http://db.csail.mit.edu/projects/cstore/)
2. MonetDB (https://www.monetdb.org/)
3. Apache Cassandra (http://cassandra.apache.org/)

O que é BigData?

- On-line transactions
- Documentos
- Emails
- Blogs
- Social Media



- Estruturado
- Semi-estruturado
- não estruturado

- Transações on-line (anytime/anywhere)
- Streaming

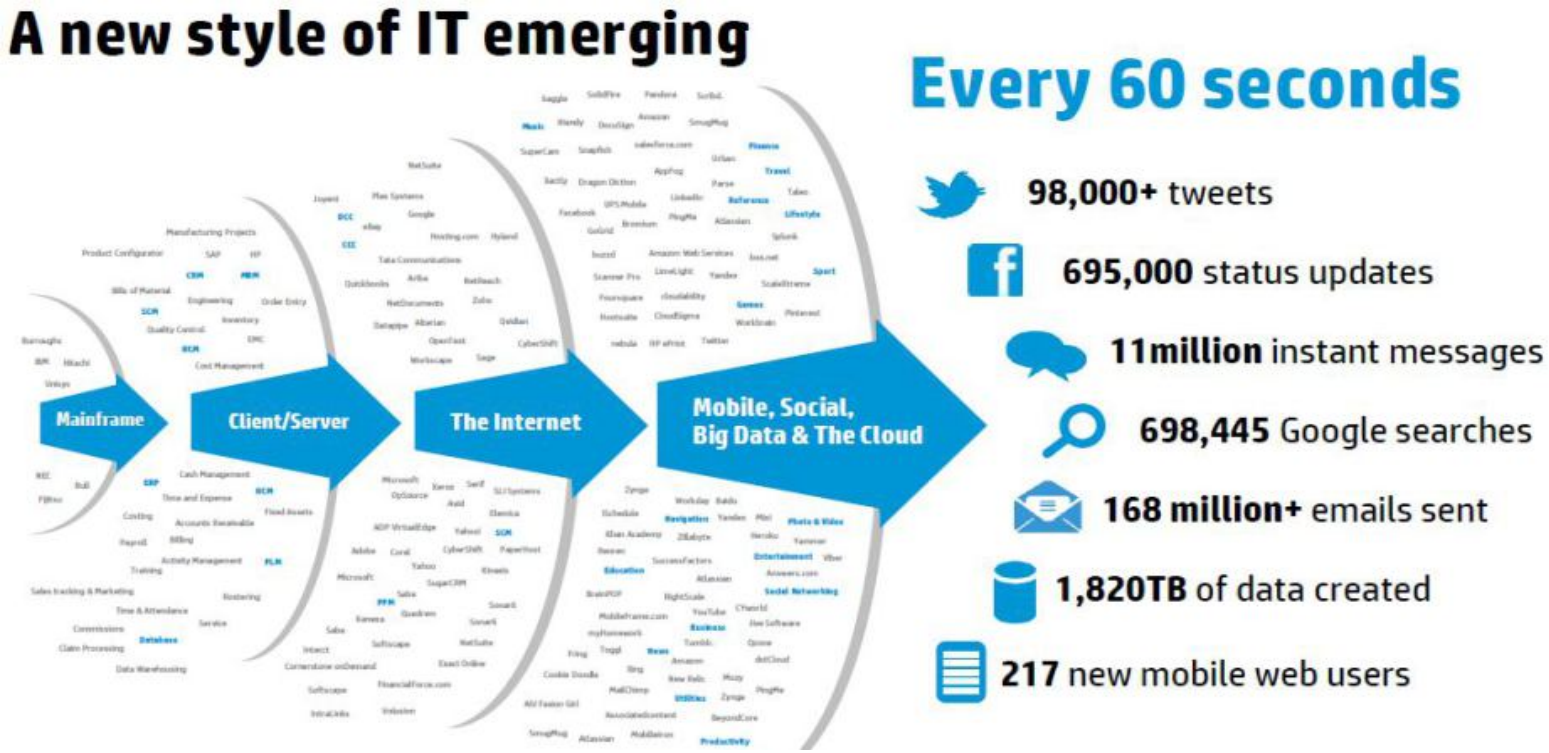
- Structured
- Semi-structured
- Unstructured

Velocidade

"Velocity as a direct consequence of the rate at which data is being collected and continuously made available."
[Dong, VLDB 2013]

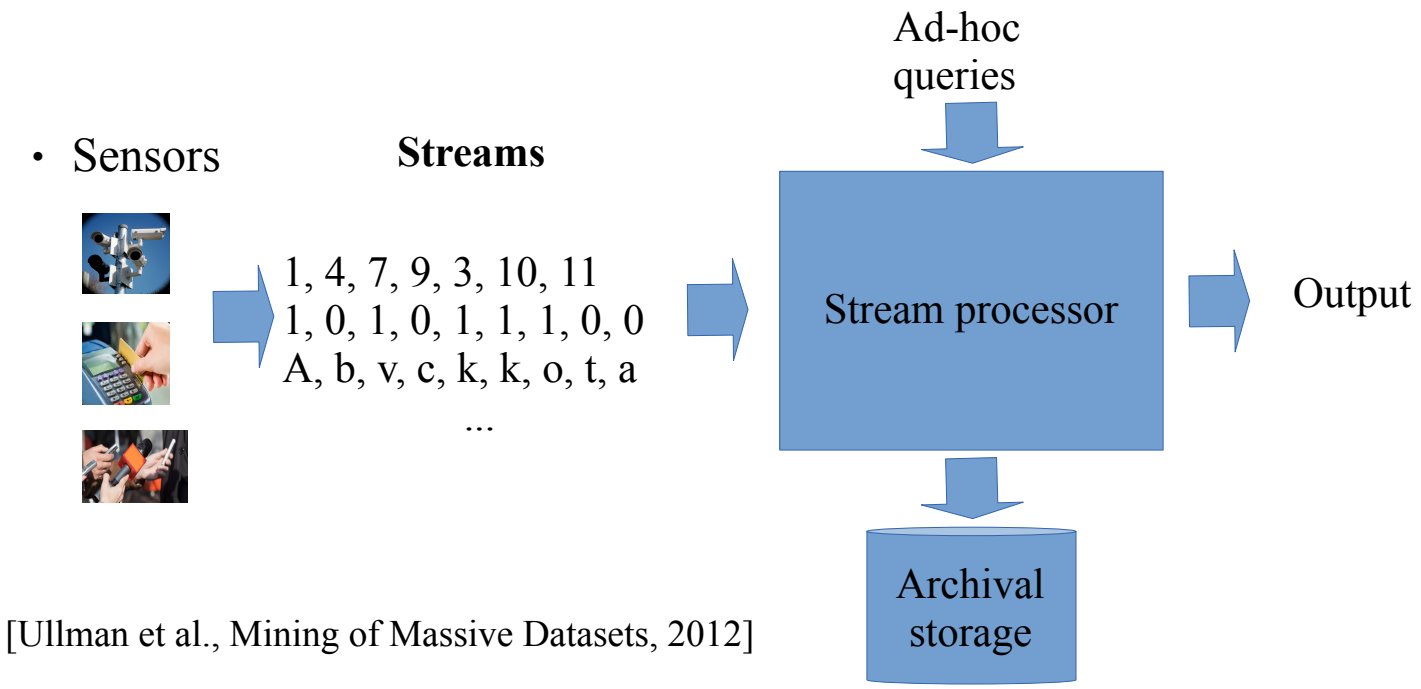
“The technology of data streaming has been investigated for several years to handle high velocity. However, the capacity of the existing streaming systems is still limited, especially when dealing with the in- creasing volume of incoming data in today’s sensor networks, telecommunication system, etc. “
[Chen, Frontiers of CS, 2013]

Velocidade



[Ravi Kalakota, 2013]

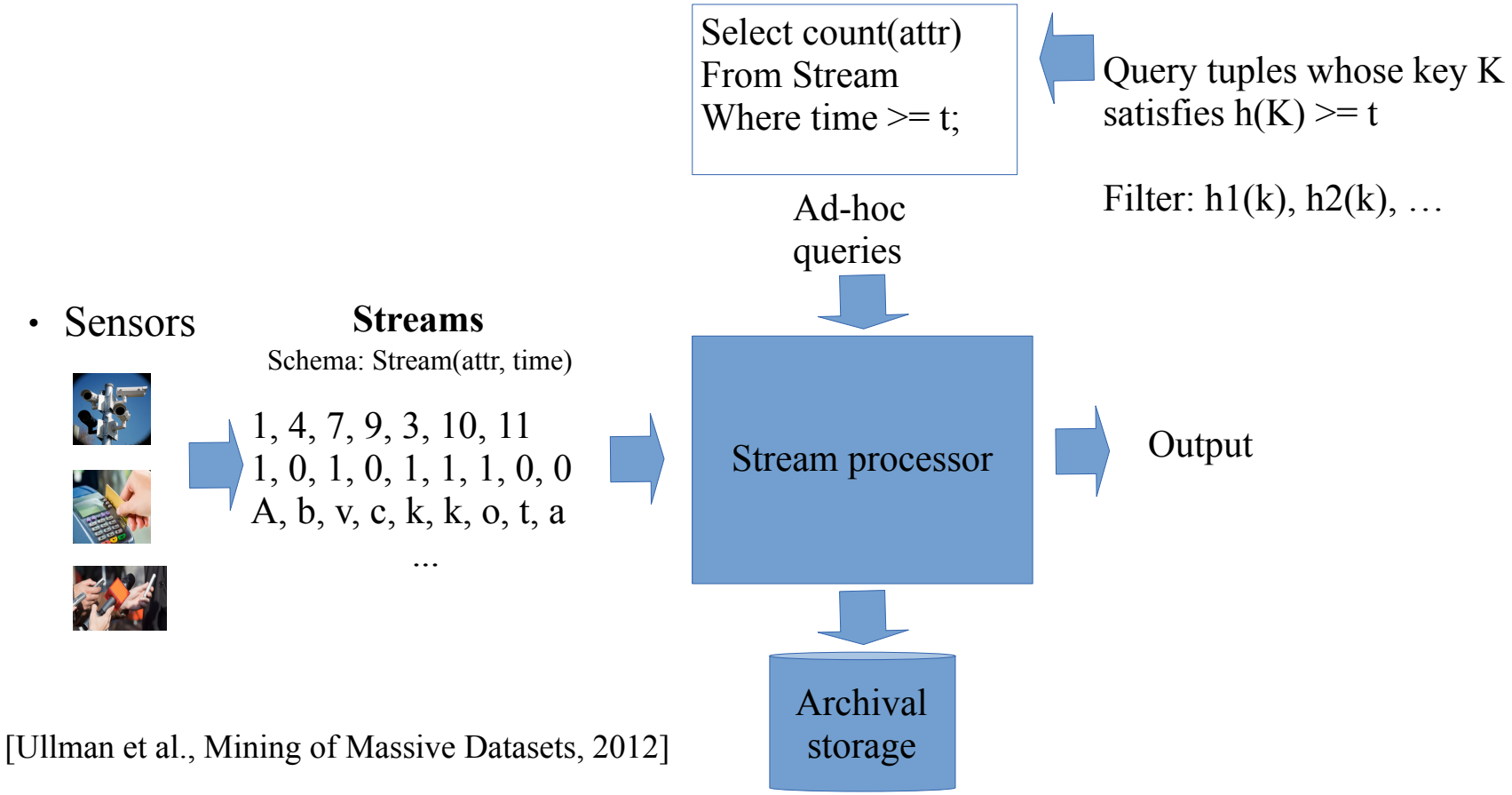
Data Streams



[Ullman et al., Mining of Massive Datasets, 2012]

Bloom Filter

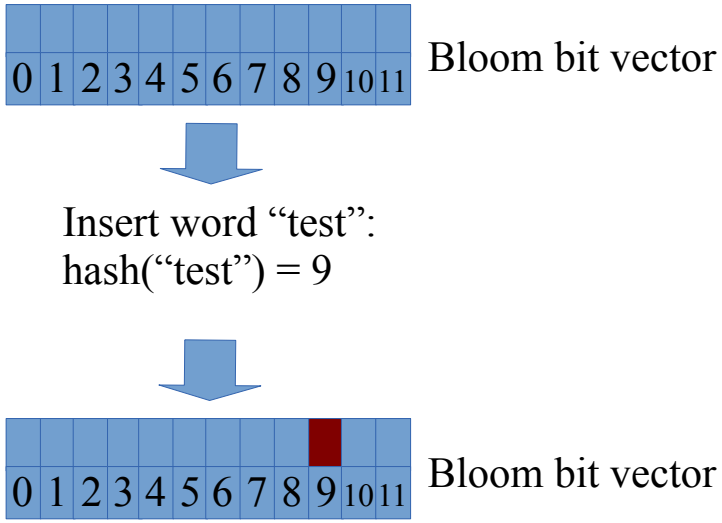
Definição: estrutura de dados probabilistica para encontrar se um elemento esta provavelmente presente no conjunto de dados.



[Ullman et al., Mining of Massive Datasets, 2012]

Bloom Filter

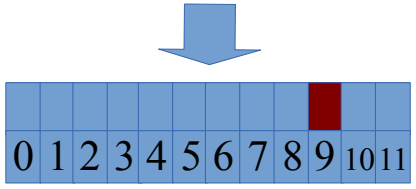
Bloom filter: valida se um elemento não está ou se provavelmente esta no conjunto de dados



Bloom Filter Example

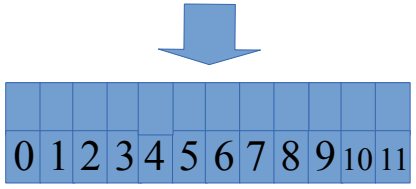
Bloom filter: valida se um elemento não está ou se provavelmente esta no conjunto de dados

encontre “test”:
hash(“test”) = 9



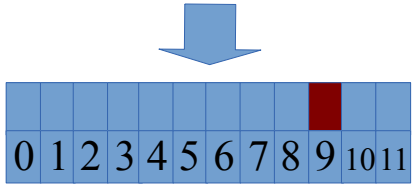
maybe

encontre “ICSE”:
hash(“ICSE”) = 4



no

encontre “ICST”:
hash(“ICST”) = 9



maybe

Interactive bloom filter example: http://
billmill.org/bloomfilter-tutorial/

Exercício

- suponha os conjuntos S={10, 32, 54} e S'={10, 31, 78}
- suponha a função h(k) = k mod m
- suponha um vetor de bits de tamanho m=8

- Considere S como um stream de tamanho m, calcule o hash de cada elemento usando a função dada acima.
- Encontre os elementos do conjunto S' no stream S

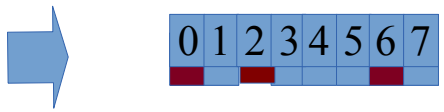
Exercício

- suponha os conjuntos S={10, 32, 54} e S'={10, 31, 78}
- suponha a função h(k) = k mod m
- suponha um vetor de bits de tamanho m=8

- Considere S como um stream de tamanho m, calcule o hash de cada elemento usando a função dada acima.
- Encontre os elementos do conjunto S' no stream S

S

h(10)=2
h(32)=0
h(54)=6



S'

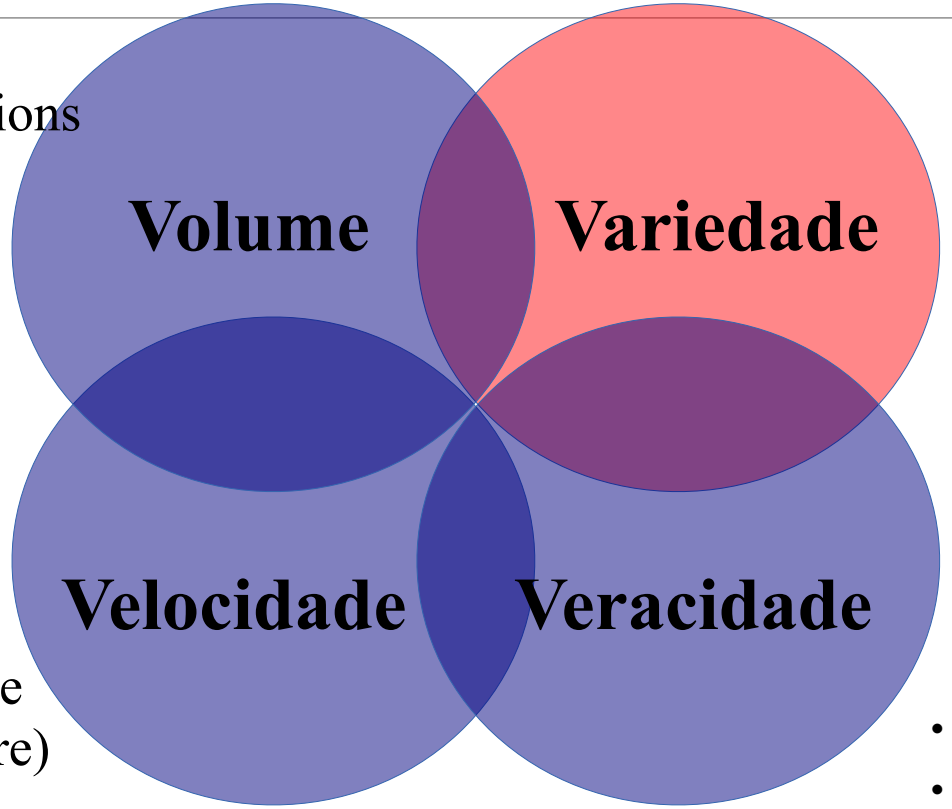
h(10)=2 → talvez
h(31)=7 → não
h(78)=2 → talvez

Sistemas para Velocidade

1.- PipelineDB (https://www.pipelinedb.com)
2.- RethinkDB (https://www.rethinkdb.com)
3.- StreamBase (http://www.streambase.com/)

O que é BigData?

- On-line transactions
- Documentos
- Emails
- Blogs
- Social Media



- Estruturado
- Semi-estruturado
- não estruturado

- Transações on-line (anytime/anywhere)
- Streaming

- Structured
- Semi-structured
- Unstructured

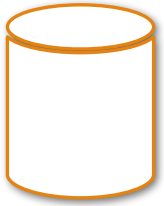
Variedade

Dados heterogêneos:
- Transactional data (business, phone calls)
- Scientific data (time series, XML, images, provenance)
- Text data (webpages)
- Graph data (social networks, RDF)

".. a mix of structured, semi-structured and unstructured data"
[Jiang, VLDB 2014]

Dados estruturados vs Não estruturados

Estruturados

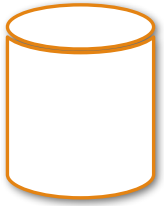


Atributo	Tipo
CPF	Varchar(11)
Nome	Varchar(50)
Endereço	Varchar(200)
Idade	Data

Não estruturados

Dados estruturados vs Não estruturados

Estruturados



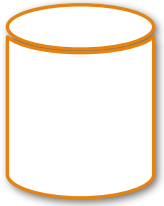
Atributo	Tipo
CPF	Varchar(11)
Nome	Varchar(50)
Endereço	Varchar(200)
Idade	Data

CPF	Nome	Endereço	Idade
123	Jose	Rua XV	10/10/2000
234	Maria	Av. Mal	01/01/1980
345	Rosana	Rua Linda	01/02/2003
...	...	...	...

Não estruturados

Dados estruturados vs Não estruturados

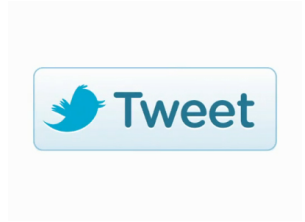
Estruturados

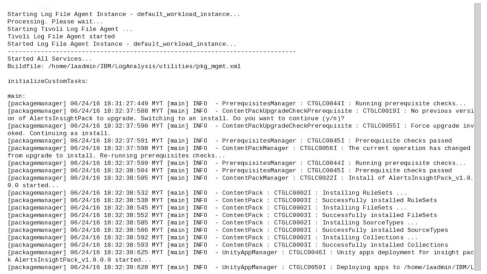


Atributo	Tipo
CPF	Varchar(11)
Nome	Varchar(50)
Endereço	Varchar(200)
Idade	Data

CPF	Nome	Endereço	Idade
123	Jose	Rua XV	10/10/2000
234	Maria	Av. Mal	01/01/1980
345	Rosana	Rua Linda	01/02/2003
...	...	...	...

Não estruturados





Variedade

"Handling diverse and messy data requires a lot of cleanup and preparation. (...) This forms 80% of the work … " [Dumbill, Forbes, 2014]

Abordagem de Data warehousing:
- Extrair das fontes de dados
- Transformar para representação única
- Carregar no BD

} 80%

- Ler e aprender com os dados :-(

Metadados importam (… bastante!)

Metadados são conjuntos de dados que descrevem e informam sobre outros dados [Oxford dictionary].

Benefícios:
- Localizar dados
- Analisar dados
- Combinar dados

Metadados em BD relacional

Metadados em BD relacional

Atributo	Tipo
CPF	Varchar(11)
Nome	Varchar(50)
Endereço	Varchar(200)
Idade	Data

Metadados em documentos (JSON)

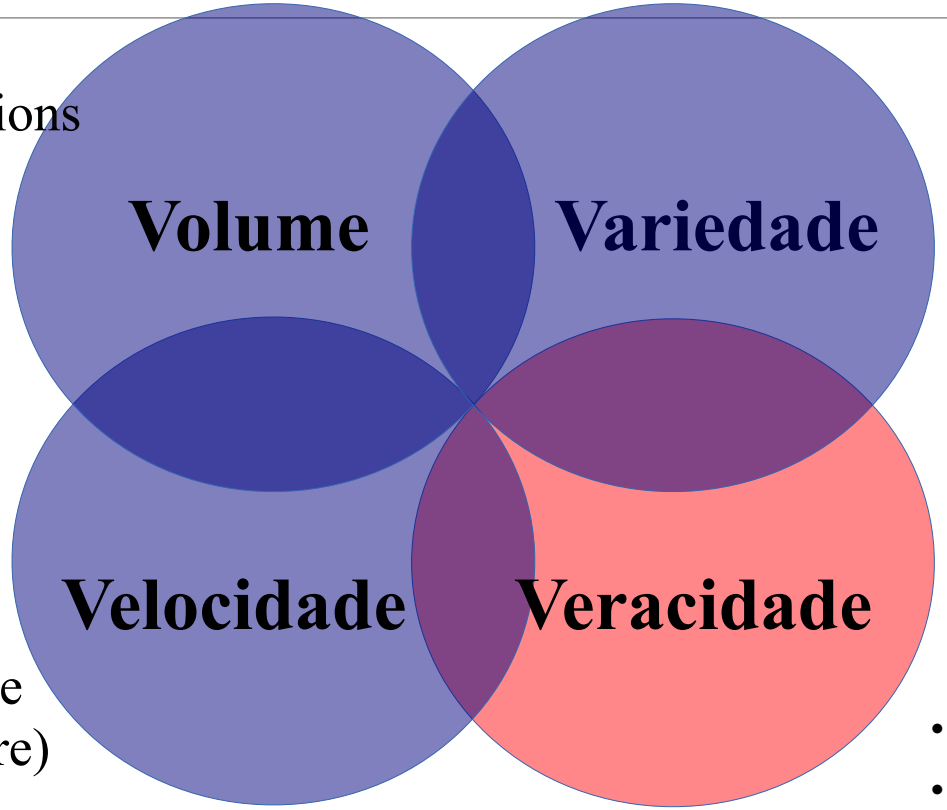
```
{
   “CPF": 104,
   "Nome": “José da Silva”,
   "Endereço": {
      "Numero": 100,
      "Rua": “Francisco H. Santos”,
      "CEP": 81531-980
   }
   "Idade": 10/01/2000,
}
```

Sistemas para Variedade

- Integração:
  - Data warehouse (Greenplum)
  - Federção (Mariposa)
  - Polystore (BigDawg)
- Schemaless:
  - NoSQL (MongoDB, Google Datastore)
  - Data lake (Hadoop)

O que é BigData?

- On-line transactions
- Documentos
- Emails
- Blogs
- Social Media



- Estruturado
- Semi-estruturado
- não estruturado

- Transações on-line (anytime/anywhere)
- Streaming

- Structured
- Semi-structured
- Unstructured

o dado em dúvida ...

- Qualidade pobre
- Amostragens ruins
- Ambiguidades
- Dados incompletos
- …

Prejuízo:
US$ 3 trilhões/ano

Harvard Business Review, 2016

Onde esta esse prejuízo?

- Satisfação do cliente (reputação da empresa)
- Multas com agências regulatórias
- Ineficiência nos processos
- Ineficiência nas tomadas de decisão
- …

http://download.101com.com/pub/tdwi/Files/DQReport.pdf

Limitar coleta de dados

- Data processing latency

"Processing latency is 24-48 hours. Standard accounts that send more than 200,000 sessions per day to Google Analytics will result in the reports being refreshed only once a day. This can delay updates to reports and metrics for up to two days. To restore intra-day processing, reduce the number of sessions you send to < 200,000 per day. For Premium accounts, this limit is extended to 2 billion hits per month."
[Google Analytics, 2014]

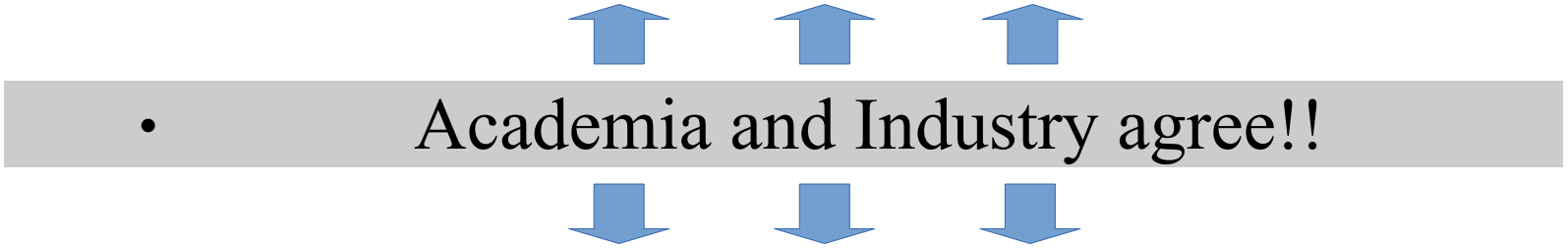
Limitar coleta de dados

- Security

"Personal data collection is a prerequisite to well-tailored services, which are in the interest of both service provider and applicant. A classical way to collect such data is to issue application forms. When considering privacy from the applicant's point of view it is unquestionable that the personal information harvested in these forms must be reduced to a minimum necessary to make the correct decision."
[N. Anciaux et al., PST 2012]

Amostragem

"Samples and synopses can lose the "long tail" in a data set, and that is increasingly where the competition for effectiveness lies." [J. Hellerstein et al., VLDB 2009]

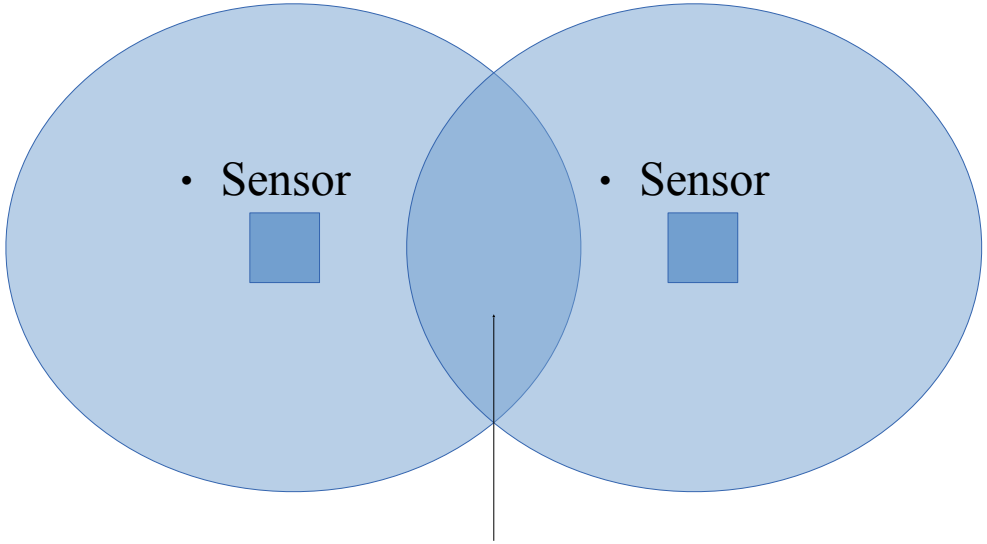


"Big data analytics and the end of sampling as we know it ... needle-in-a-haystack problems don't lend themselves well to samples".

ComputerWeekly.com

- Eliminar redundancia
- Eliminar “Cold Spots”
- Limpar dados

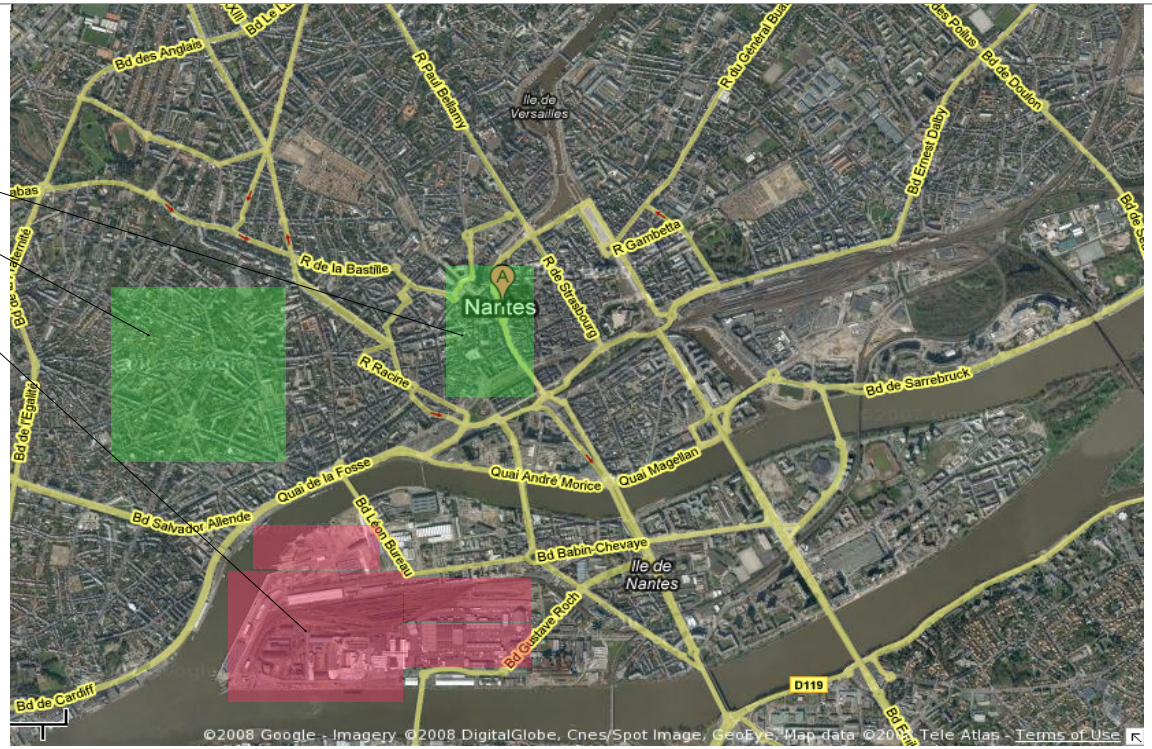
Eliminar redundâncias



- Armazena e processa 2x?
  - (eg. Dados de Video surveillance)
- Quem eliminar?
  - (eg. dado de temperatura -> um sensor é necessário)

Eliminando “cold spots”

- Busca somente os dados necessários e o resto descarrega em FITA ou apaga!!



- Detectar “cold spots” com politicas de monitoramento e limpeza de dados [Basin et al. TSE, 2013]
- (eg. somente dados de 2018!!)

Exemplo: quantas matriculas em disciplinas tem Maria? **(1)**

aluno	instructor
Maria	Eduardo
Maria	André
Maria	David

Tabela: T1

aluno	disciplina
Maria	ci218
Maria	ci056
Maria	ci057

Tabela: T2

Exemplo: quantas matriculas em disciplinas tem Maria?

programa=> SELECT * from T1 natural join T2;

aluno	instrutor	disciplina
Maria	Eduardo	ci218
Maria	Eduardo	ci056
Maria	Eduardo	ci057
Maria	André	ci218
Maria	André	ci056
Maria	André	ci057
Maria	David	ci218
Maria	David	ci056
Maria	David	ci057

9 matriculas!!!

Agenda do Curso

- Modelos de Armazenamento
- Modelos de Indexação
- Modelagem de dados: abstrato e lógico
- Processamento de consultas (Simone)
- Processamento distribuído de dados (Ramiro)

BigData

A Bullet for The General (1967)

Aula #1 - Processamento de BigData

EDUARDO CUNHA DE ALMEIDA