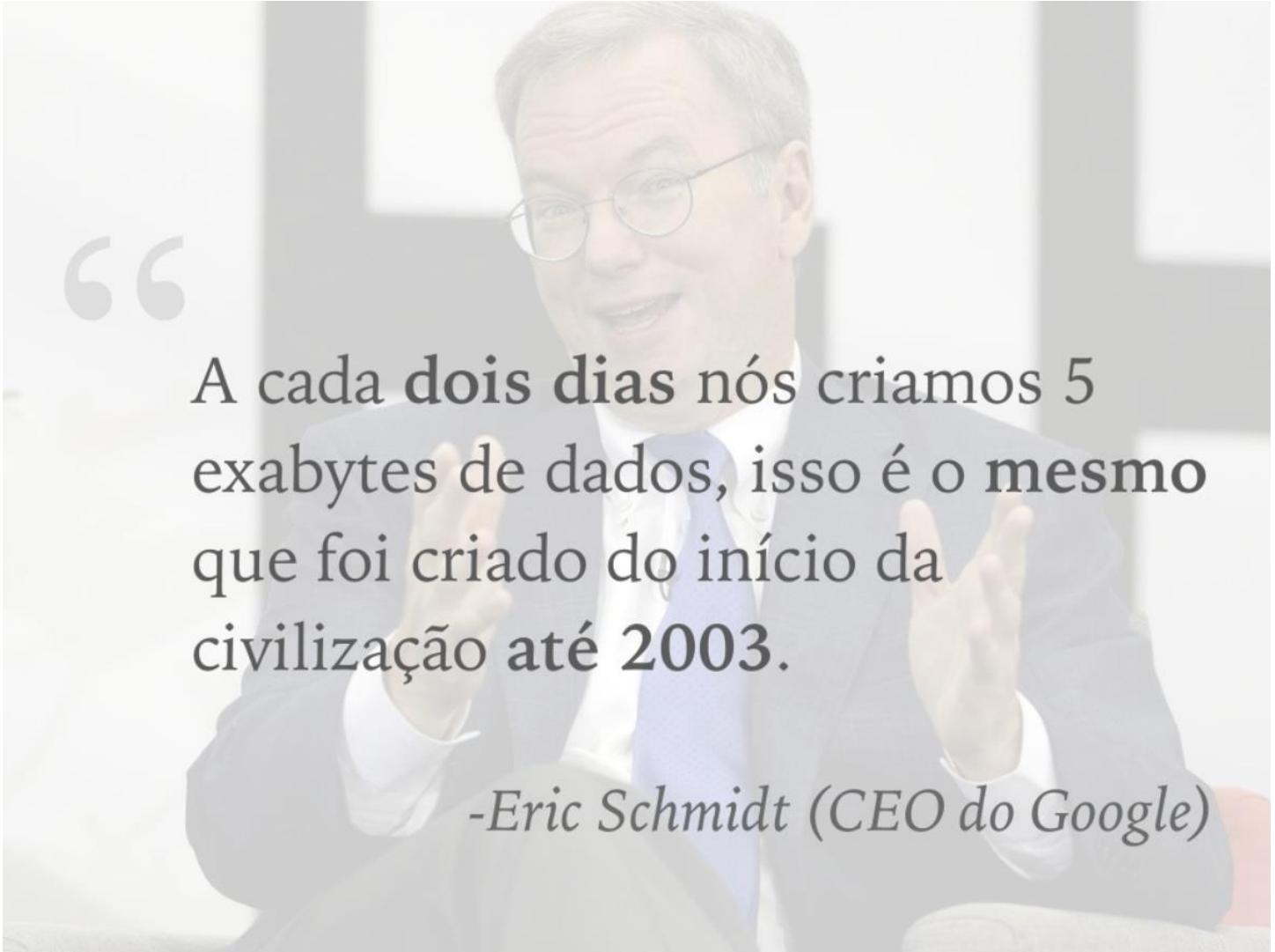

FIAP

INTRODUÇÃO A
BIG DATA


“
A cada **dois dias** nós criamos 5 exabytes de dados, isso é o **mesmo** que foi criado do início da civilização **até 2003**.
-Eric Schmidt (CEO do Google)

O mundo armazenará 200 Zettabytes de dados até 2025

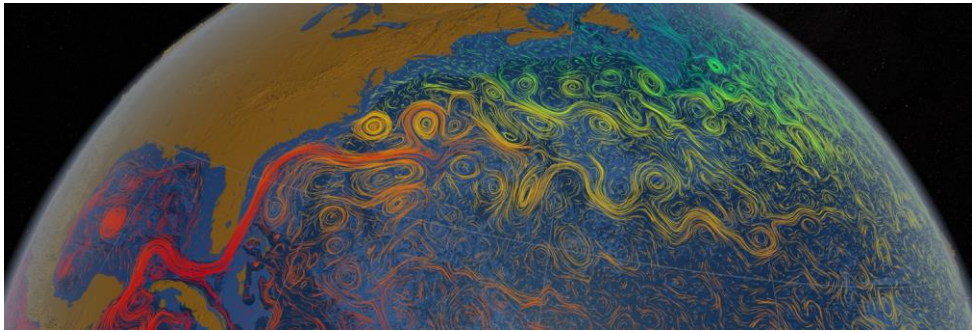
Fonte: Cybersecurity Ventures

Visualizando o volumes de dados

Byte	1 grão de arroz
Kilobyte	1 xícara de arroz
Megabyte	8 sacos de arroz
Gigabyte	1 container de arroz
Terabyte	2 navios cargueiros
Petabyte	Suficiente para cobrir a cidade de Campinas.
Exabyte	Suficiente para cobrir os estados de Minas Gerais, Rio de Janeiro, Espírito Santo e São Paulo.
Zettabyte	Preenche o oceano Pacífico.

Big Data não é novidade!

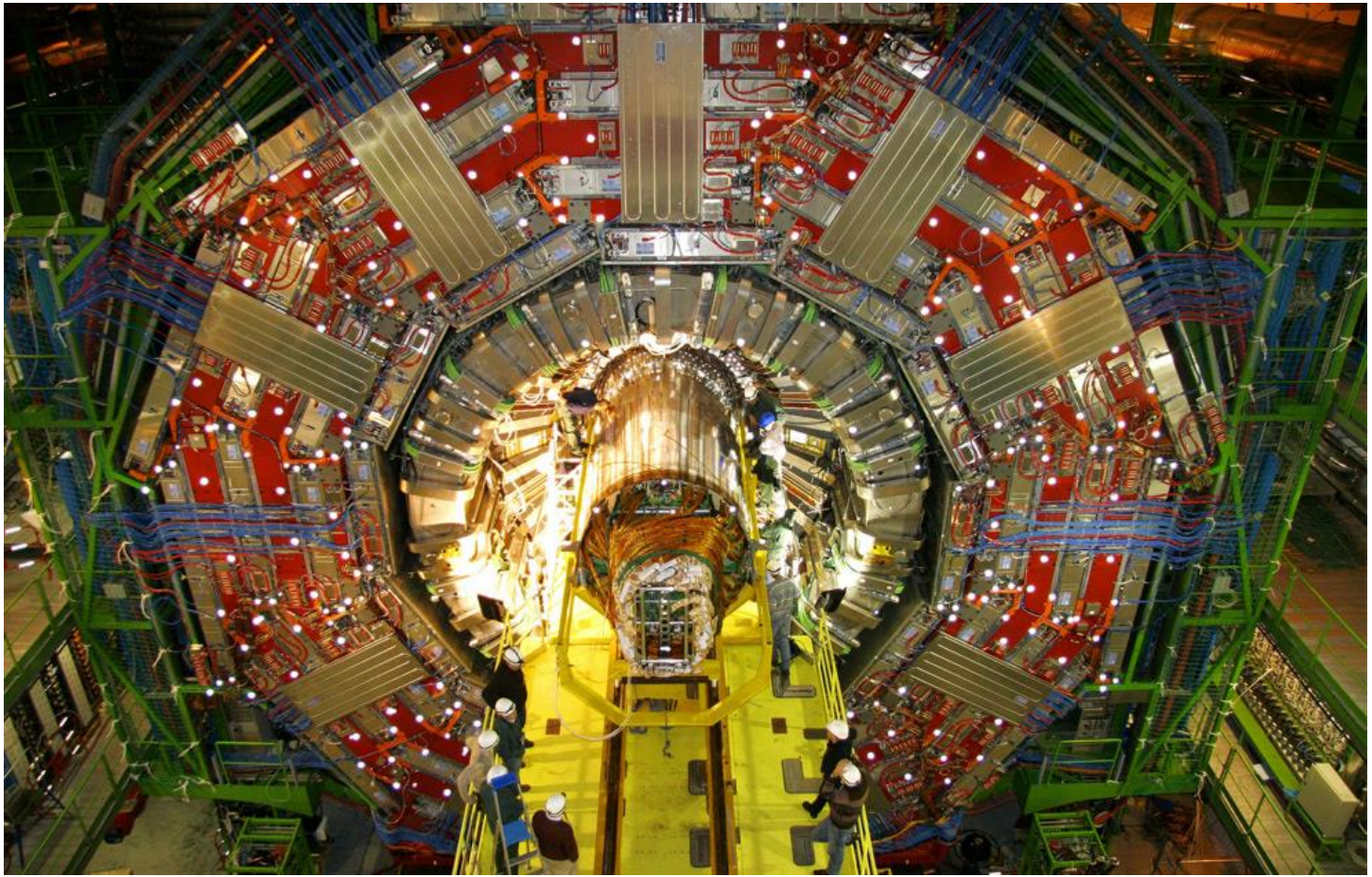
- Há décadas temos necessidade de processar grandes volumes de dados.
- Astronomia, geologia, oceanografia, meteorologia sempre trabalharam com grande quantidade de dados.



Big Data não é novidade!

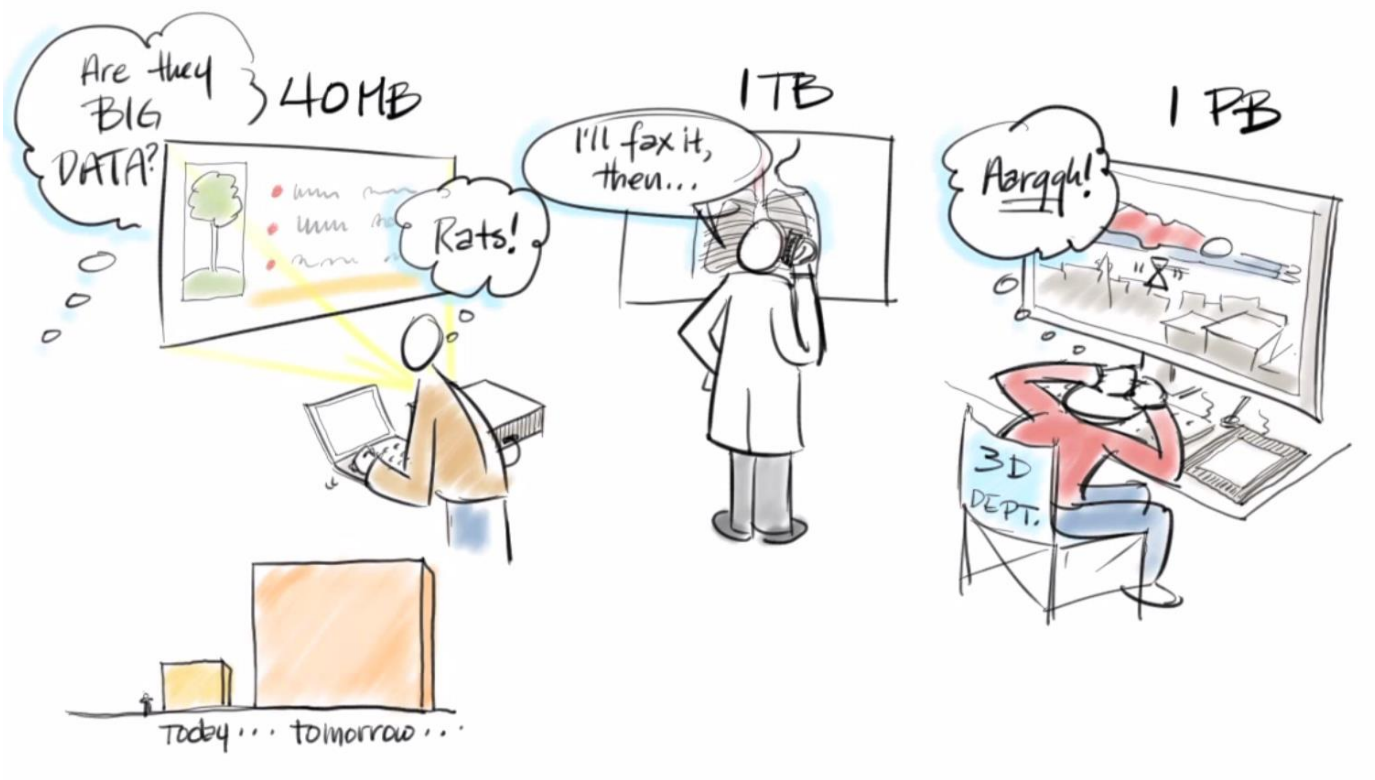

```
EDIT       SYSADM.DEMO.JCLLIB(#IMSESDS) - 01.01            Columns 00001 00072
Command ===>                                                Scroll ===> CSR
****** ***************************** Top of Data ********************************
000001 //SYSADMA  JOB A123,'BIN-7 QUASAR',CLASS=A,MSGCLASS=Y,NOTIFY=&SYSUID
000002 //*--------------------------------------------------------------------
000003 //*          RUN THIS JOB TO CREATE IMS V8 DATASETS                    |
000004 //*  AUTHOR : QUASAR CHUNAWALA                  DATE : 11/06/2010      |
000005 //*--------------------------------------------------------------------
000006 //STEP001 EXEC PGM=IDCAMS
000007 //SYSPRINT  DD SYSOUT=*
000008 //SYSIN     DD *
000009       DEFINE CLUSTER(NAME(IMS810.INVENDB)                              -
000010       TRACKS(1,1)                                                      -
000011       NONINDEXED                                                       -
000012       RECORDSIZE(2041,2041)                                            -
000013       VOLUMES(USER01)                                                  -
000014       CONTROLINTERVALSIZE(2048))
000015 //
```

- O processamento exige *hardware* específico, *softwares* e desenvolvedores com habilidades analíticas específicas.
- Aumento pela demanda por *hardware* de baixo custo e *softwares* que pudessem ser desenvolvidos por programadores com habilidades de programação convencional.



Large Hydron Collider (LHC) da CERN gera 15 PB por ano

O que é Big Data



Fonte: EMC²

O que é *Big Data*

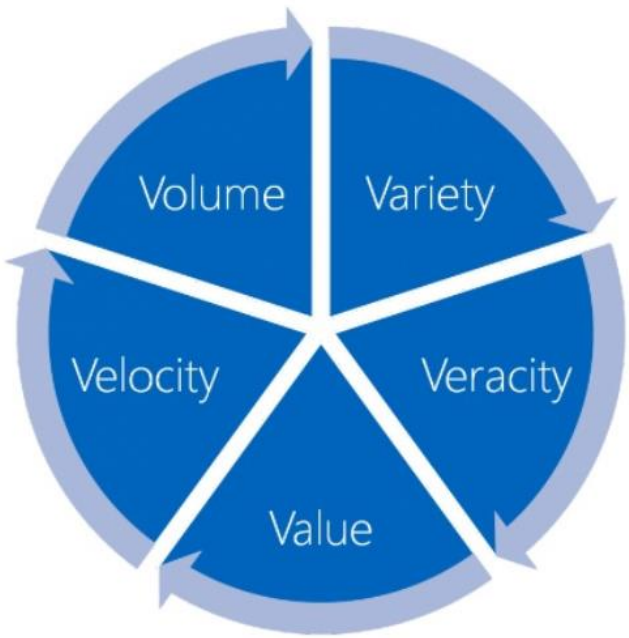
- Segundo o McKinsey Global Institute [2011], O termo *Big Data* implica no uso de metodologias e ferramentas para processamento e análise de dados que possam produzir resultados úteis que não possam ser deduzidas/calculadas, de maneira eficiente, através de outros métodos.

Fonte: EMC²

O que é *Big Data*

Fonte: EMC²

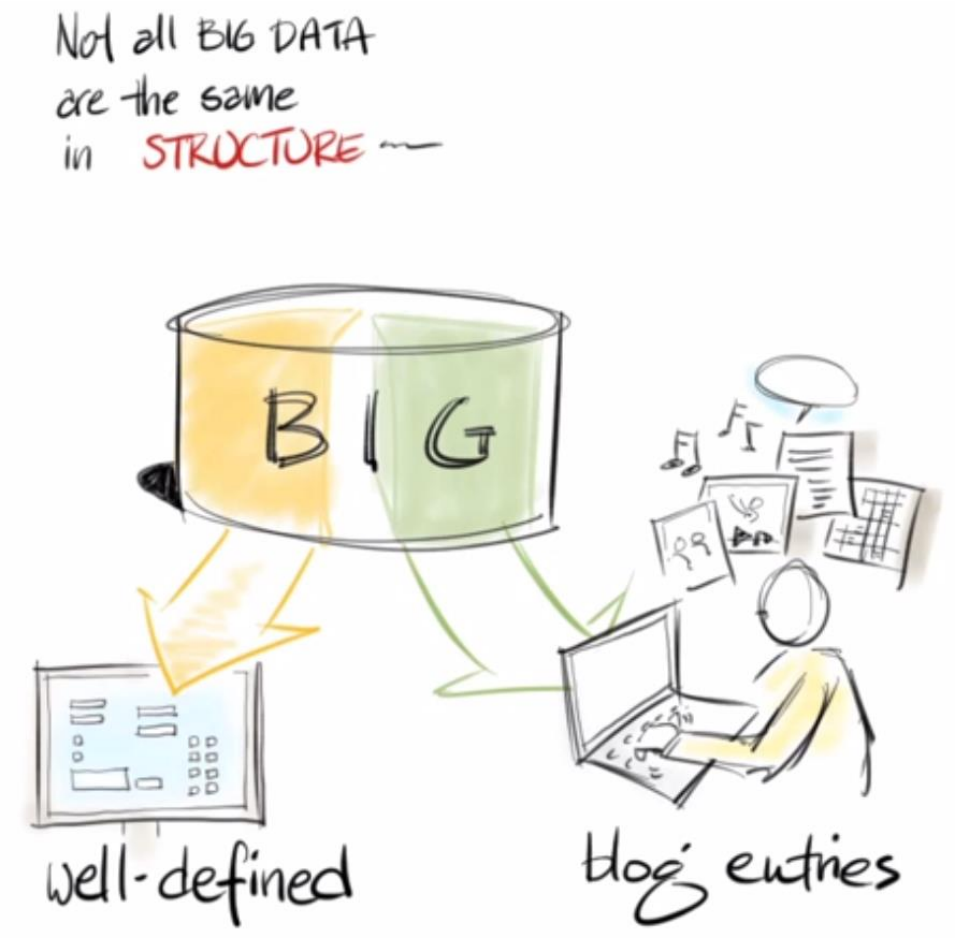
Os V's



- **Volume**: Grandes de dados, desafios de armazenamento
- **Variedade**: Diferentes formatos de dados estruturados e/ou não
- **Veracidade**: Acurácia e autenticidade
- **Valor**: Retorno desses dados para o negócio/sociedade
- **Velocidade**: Tempo entre o dado gerado e analisado

Fonte: Oracle

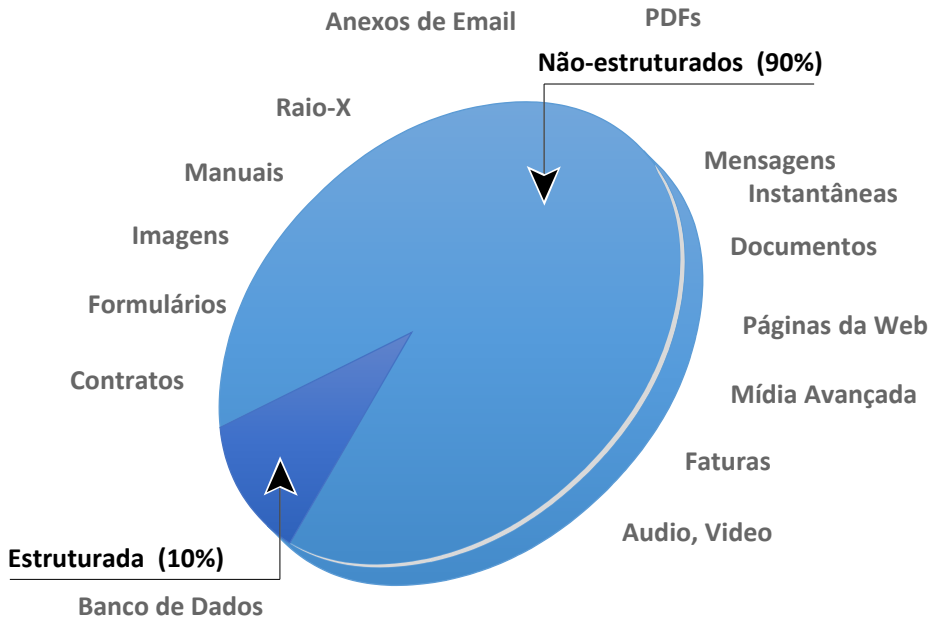
O que é *Big Data*



Fonte: EMC²

Tipos de dados

- Os dados podem ser classificados como:
  - Estruturado
  - Semi-estruturado
  - Não-estruturado
- A maioria dos dados que estão sendo criados são não-estruturados



Fonte: EMC²

Teorema CAP

- Em qualquer sistema distribuído *stateful* é preciso escolher entre
  - Consistency (consistência forte). Todos os nós veem os mesmos dados ao mesmo tempo.
  - Availability (alta disponibilidade). Toda solicitação recebe uma resposta, seja ela bem-sucedida ou não.
  - Network Partition Tolerance (tolerância a particionamento dos dados na rede). O sistema continua funcionando mesmo que mensagens sejam perdidas ou parte do sistema falhe.
- Entre as três propriedades, somente duas podem ser garantidas ao mesmo tempo.

(Eric Brewer, 2000)

http://www.cs.berkeley.edu/~brewer/cs262b-2004/PODC-keynote.pdf

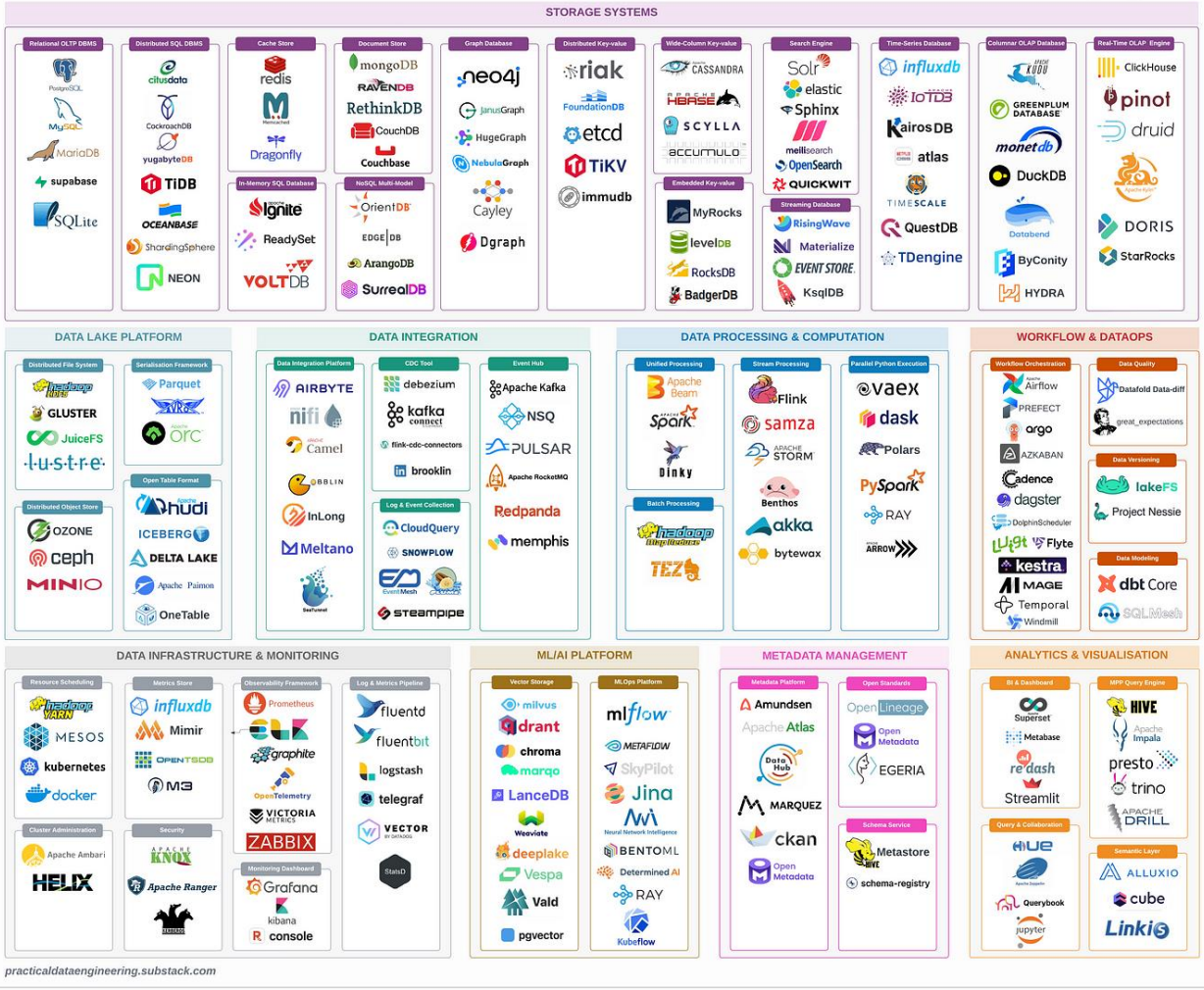
Teorema CAP

- **CAp** – Bases de dados tradicionais relacionais.
  - Em caso de falha da rede ou de grande latência, elas não consegue responder a todos os pedidos.

- **cAP** – Bases de dados NOSQL.
  - Estes são sistemas altamente tolerantes a falhas, desde que existam um grande numero de servidores suportando o sistema.
  - Fornecem disponibilidade mas consistência eventual.

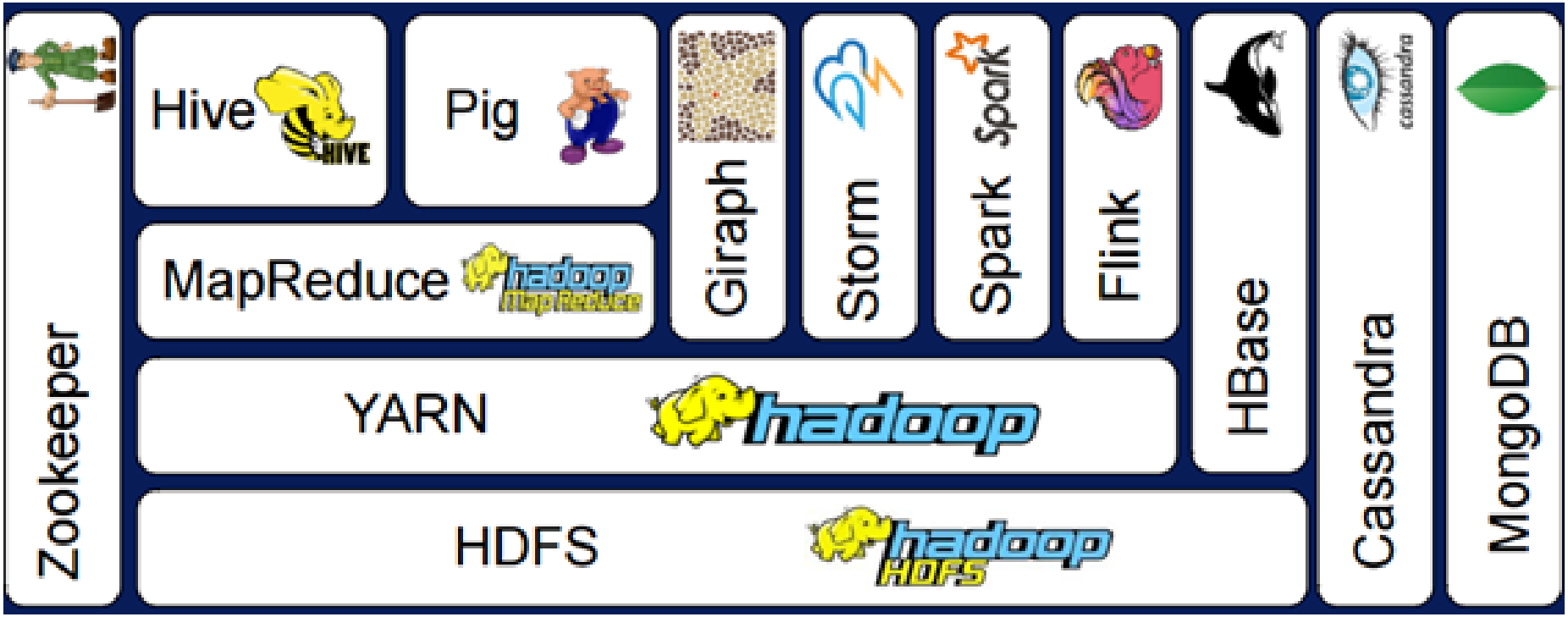
- **CaP** – O sistema não dá garantias de sempre estar disponível.
  - É um sistema onde um nó falha e os outros não podem responder as solicitações.

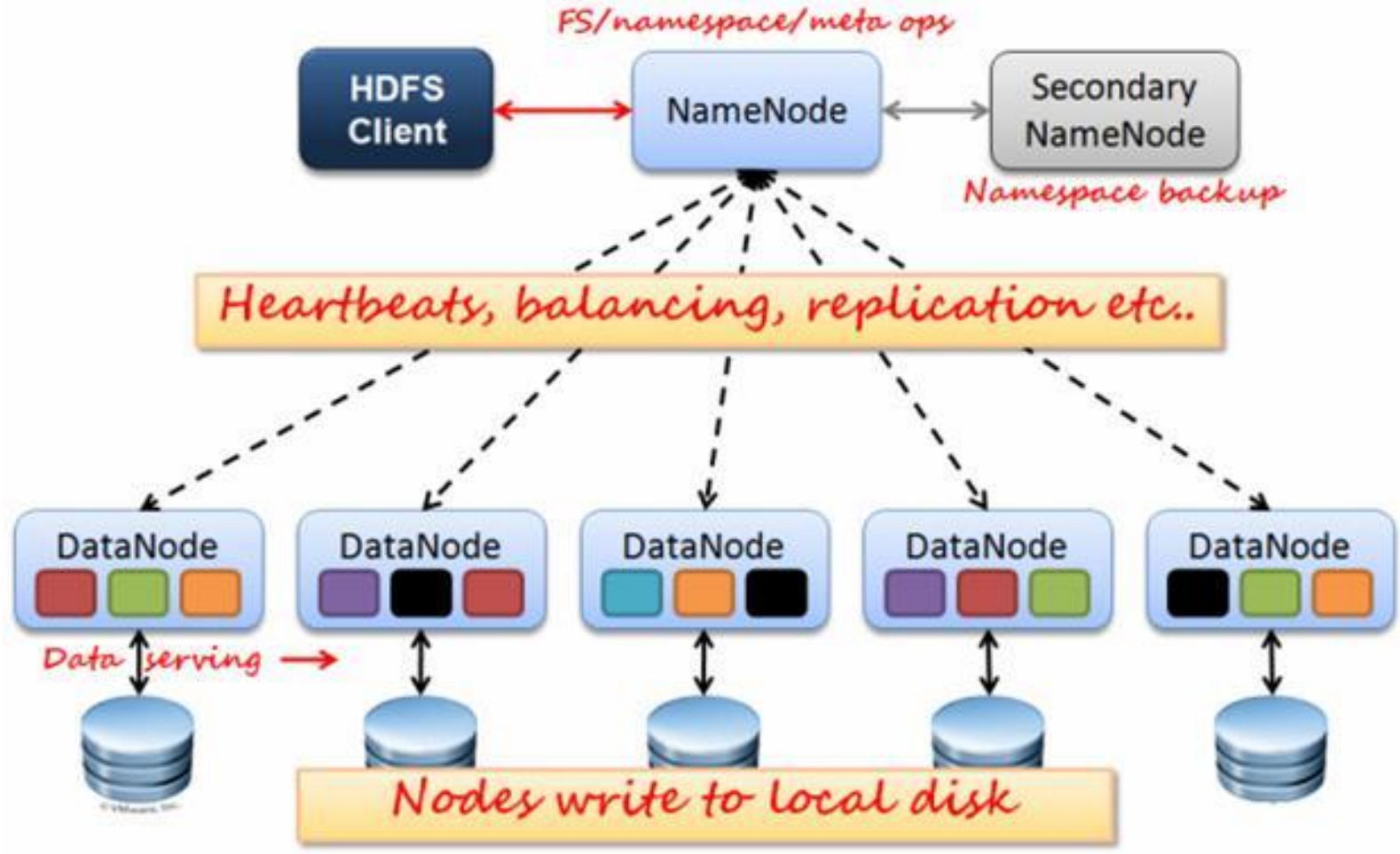
OPEN SOURCE DATA ENGINEERING LANDSCAPE 2024



practicaldataengineering.substack.com

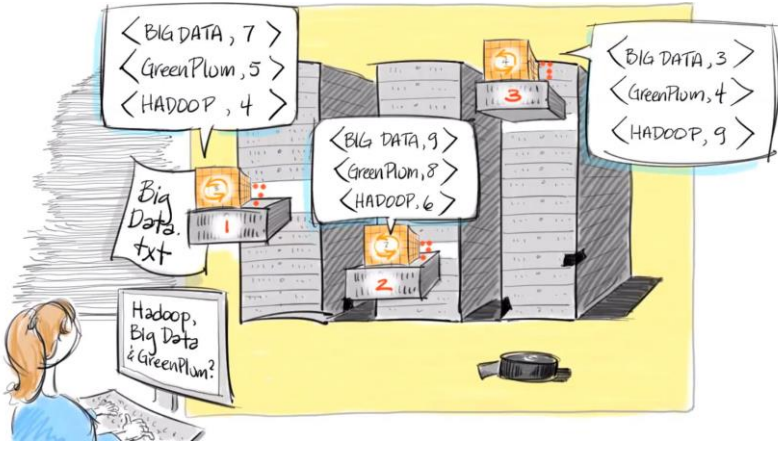
Ecossistema Big Data




FS/namespace/meta ops
HDFS Client
NameNode
Secondary NameNode
Namespace backup
Heartbeats, balancing, replication etc..
DataNode
DataNode
DataNode
DataNode
DataNode
Data serving
Nodes write to local disk

MapReduce

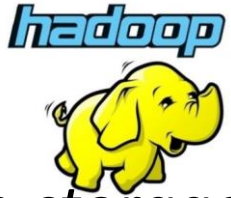
- *Map*:
  - Obtêm uma lista de pares <*Key, Value*>, processa os pares e gera um conjunto de de pares <*Key, Value*> intermediário.
  - Repassa o valor intermediário para a função Reduce.
  - Cada par é processado em paralelo.



MapReduce

- Reduce:
  - Processa todos os valores associados com a mesma <*Key*>.
  - Mescla os valores para formar um conjunto de valores possivelmente menor.
  - Geralmente, apenas um valor de saída de 0 ou 1 é produzido a cada chamada Reduce.
  - Os valores intermediários são fornecidos à função Reduce do usuário por um iterador permitindo identificar listas de valores que são grandes demais para a memória.

Hadoop



- Segundo a Hadoop, “Hadoop é um *storage* confiável **e** um sistema analítico” [2014]
- Composto por duas partes essenciais:
  - o Hadoop Distributed Filesystem (HDFS), sistema de arquivos distribuído e confiável, responsável pelo armazenamento dos dados
  - Hadoop MapReduce, responsável pela análise e processamento dos dados.
- O nome do projeto veio do elefante de pelúcia que pertencia ao filho do criador, Doug Cutting.

Pig e Pig Latin

- Pig é um mecanismo para executar os fluxos de dados de modo paralelo ao Hadoop.
- Usa uma linguagem chamada Pig Latin para expressar esses fluxos de dados.
- Com a Pig Latin, é possível descrever como os dados de uma ou mais entradas devem ser lidos, processados e, depois, armazenados em uma ou mais saídas de modo paralelo.

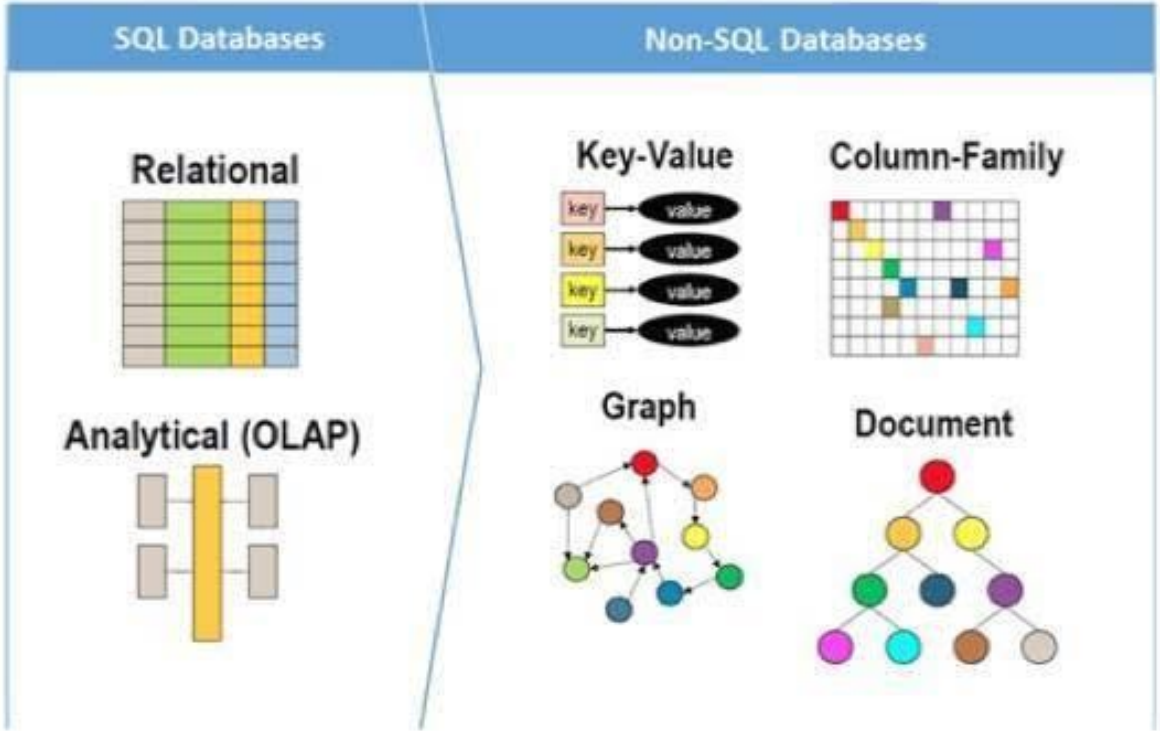
Hive e HiveQL

- Hive é *framework* para soluções de *Data Warehousing* executado no ambiente Hadoop.
- HiveQL é uma linguagem declarativa, similar ao SQL, usada para criar programas executáveis no Hive.
- O compilador HiveQL traduz os comando em HiveQL em *jobs* do MapReduce e os envia para o Hadoop executar.

418 systems in ranking, March 2024

Rank Mar 2024	Rank Feb 2024	Rank Mar 2023	DBMS	Database Model	Score Mar 2024	Score Feb 2024	Score Mar 2023
1.	1.	1.	Oracle	Relational, Multi-model	1221.06	-20.39	-40.23
2.	2.	2.	MySQL	Relational, Multi-model	1101.50	-5.17	-81.29
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	845.81	-7.76	-76.20
4.	4.	4.	PostgreSQL	Relational, Multi-model	634.91	+5.50	+21.08
5.	5.	5.	MongoDB	Document, Multi-model	424.53	+4.18	-34.25
6.	6.	6.	Redis	Key-value, Multi-model	157.00	-3.71	-15.45
7.	7.	↑ 8.	Elasticsearch	Search engine, Multi-model	134.79	-0.95	-4.28
8.	8.	↓ 7.	IBM Db2	Relational, Multi-model	127.75	-4.47	-15.17
9.	9.	↑ 11.	Snowflake	Relational	125.38	-2.07	+10.98
10.	10.	↓ 9.	SQLite	Relational	118.16	+0.88	-15.66
11.	11.	↓ 10.	Microsoft Access	Relational	107.93	-5.24	-24.13
12.	12.	12.	Cassandra	Wide column, Multi-model	104.59	-4.69	-9.20
13.	13.	13.	MariaDB	Relational, Multi-model	95.03	-2.20	-1.81
14.	14.	14.	Splunk	Search engine	89.68	-1.97	+1.71
15.	↑ 16.	↑ 16.	Microsoft Azure SQL Database	Relational, Multi-model	78.51	-1.06	+1.06
16.	↓ 15.	↓ 15.	Amazon DynamoDB	Multi-model	77.72	-5.18	-3.05
17.	17.	↑ 19.	Databricks	Multi-model	74.34	-2.57	+13.48
18.	18.	↓ 17.	Hive	Relational	64.82	-0.99	-6.09

noSQL: “Not Only SQL”



OBRIGADO

FIAP


FIAP